Journal of Applied Research on Industrial Engineering

www.journal-aprie.com

J. Appl. Res. Ind. Eng. Vol. 11, No. 4 (2024) 669–705.

Paper Type: Review Paper

A Review of Methods for Detecting Multimodal Emotions in Sound, Image and Text

Seyed Sadegh Hosseini[1], Mohammadreza Yamaghani[1,*], Soodabeh Poorzaker Arabani[1]

[1] Department of Computer Engineering, Lahijan Branch, Islamic Azad University, Lahijan, Iran; sadeghhosseini@malayeru.ac.ir; o_yamaghani@iau.ac.ir; arabani@liau.ac.ir.

Citation:

Received: 26 August 2023 *Revised: 20 October 2023* *Accepted: 15 December 2023*	Hosseini, S. S., Yamaghani, M., & Poorzaker Arabani, S. (2024). A review of methods for detecting multidimensional emotions in sound, image and text. *Journal of applied research on industrial engineering, 11*(4), 669-705.

Abstract

Emotional computing synergizes the understanding and quantification of emotions, drawing on diverse data sources such as text, audio, and visual indicators. A challenge arises when attempting to discern authentic emotions from those concealed deliberately via facial cues, vocal nuances, and other communicative behaviors. Integrating multiple physiological and behavioural signals can give more profound insights into an individual's emotional state. Historically, research has predominantly concentrated on a singular facet of emotional computing. In contrast, our study offers an in-depth exploration of its pivotal domains, encompassing emotional models, Databases (DBs), and contemporary developments. We begin by elucidating two prevalent emotional models and then examine a renowned sentiment analysis DB. Subsequently, we delve into cutting-edge emotion detection and analysis methodologies across varied sensory channels, elaborating on their design and operational principles. In conclusion, the fundamental principles of emotional computing and its real-world implications are discussed. This review endeavors to provide researchers from academia and industry with a holistic understanding of the latest progress in this domain.

Keywords: Emotion recognition, Machine learning, Deep learning, Multimodal emotions, Dataset, Information fusion, Feature extraction.

1 | Introduction

Emotional computing, a term encompassing the recognition and analysis of human emotions [1], was introduced by Professor Picard in 1997 [2]. This concept emphasizes the computer's ability to understand,

Corresponding Author: o_yamaghani@iau.ac.ir
https://doi.org/10.22105/jarie.2024.413341.1560

interpret, and react intelligently to human emotions [2], [3]. Applied research is increasingly eager to develop cognitive and intelligent systems that recognize individual emotions and swiftly generate apt responses [4]–[6]. For example, smart vehicle technology may leverage real-time emotional state monitoring of the driver to mitigate potential accidents [7]. It can also be used in sports predictions [8]. Moreover, emotional computing is pivotal in gauging sentiments across various social media platforms [9]. This has led many scholars [3], [10] to view emotional computing as instrumental in advancing human-centric artificial intelligence. The domain of emotional computing primarily bifurcates into Emotion Recognition (ER) and emotion analysis [11]–[14]. Psychological research highlights two prevalent theories for human emotion modeling: the discrete (or categorical) emotion model [15] and the dimensional emotion model [16], both essential for emotion quantification. ER targets the identification of specific human emotional states, including both primary and subsequent emotions [17]. Methods have been proposed for Visual Emotion Recognition (VER) [18], Audio Emotion Recognition/Speech Emotion Recognition (AER/SER) [19], and Physiological Emotion Recognition (PER) [20]. Conversely, sentiment analysis evaluates opinions, typically classifying them as positive, negative, or neutral, especially within social and marketing contexts [21], [22]. Notably, a context-level attention framework was created to determine the overall emotion and detect specific ones within speech [23]. The public release of diverse Databases (DBs) has accelerated advancements in emotional computing, emphasizing both unimodal and multimodal DBs. Sentiment computing systems emerging from this development prominently incorporate Machine Learning (ML) and Deep Learning (DL) methodologies. Research indicates that in daily interactions, human emotions are mainly conveyed through facial expressions (55%), vocal nuances (38%), and verbal language (7%) [24]. For the context of this review, we categorize text, audio, and video signals as non-contact data sources. The vast data pool from uninhibited expressions on digital platforms allows researchers to pinpoint even nuanced emotions [25]–[27]. ER devoid of physical touch may not be wholly accurate as individuals can mask their genuine emotions [20]. Physiological signals like EEG and ECG offer more consistent real-time emotion predictions as these parameters are challenging to manipulate intentionally [28], [29]. Considering the intricate nature of human emotions [30], current research endeavors to discern multimodal emotions without physical intervention [31]. The synergy of text, audio, and visual modalities delivers richer information than individual sources [32], [33], mirroring our brain's multi-sensory data interpretation process. Advances in wearable sensors have facilitated ER based on multiple physiological signals. Combining non-contact and physiological methods can yield a comprehensive emotional computing framework, capturing even nuanced emotions [34], [35]. Choosing apt unimodal sentiment data and effective multimodal fusion techniques [36] is crucial for effective sentiment analysis systems [37]. In general, the identification of emotions includes the following steps (*Fig. 1*). Herein, we offer a concise overview of multidimensional ER techniques (*Fig. 2*).

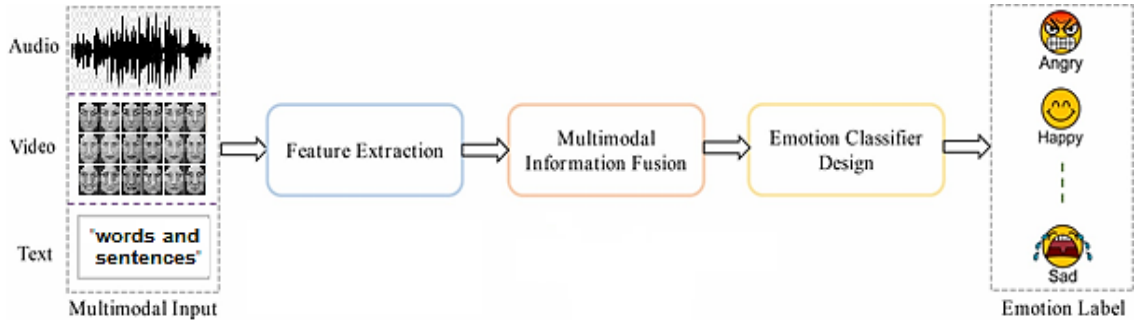


Fig. 1. The workflow of a Multimodal Emotion Recognition (MER) system comprises three primary steps: 1) extraction of multimodal features, 2) fusion of multimodal information, and 3) design of the emotion classifier.

Although abundant literature exists on emotion computing, many studies focus on non-physically invasive ER. Earlier works have touched upon Facial Expression Recognition (FER) [38]–[41], ML applications in sentiment analysis [42], and DL approaches for SER [43].

However, limited comprehensive reviews exist on physiology-based ER and physical and physiological data fusion.

Most extant reviews adopt a niche focus, occasionally bypassing DL-based sentiment recognition and multimodal sentiment analysis developments. Many lack insights into the efficacy of contemporary methods in emotion detection. This paper seeks to fill these gaps, presenting an all-encompassing perspective on emotional computing, discussing various methods findings, and positing future research trajectories. Specifically, we confine our exploration to multimodal emotional computations and their classification techniques. Our review encompasses scholarly articles from conferences and journals, as represented in *Fig. 2*, facilitating the reader's journey through this intricate research domain. We provide an exhaustive classification of state-of-the-art Sentiment Computing methods, highlighting both ML and DL techniques and their applicability. We introduce DBs that are pivotal for sentiment computing and offer a comparative analysis of selected methodologies. Finally, we delve into the multifaceted impacts and practical applications of emotional computing, charting potential future research avenues in emotion identification and analysis.

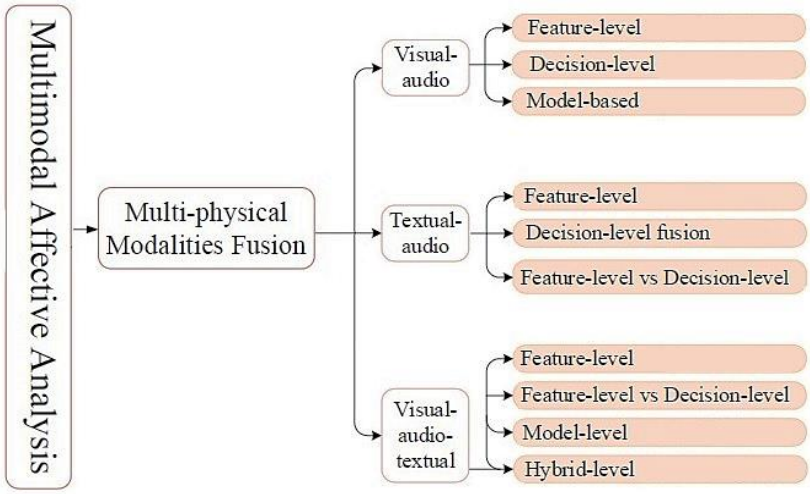


Fig. 2. This paper provides a classification of non-contact emotional calculations based on existing methods involving audio, text, and image combinations.

The structure of this article is as follows:

Section 1 provides an introduction. Section 2 reviews the literature, highlights relevant studies and outlines the scope of the field. Section 3 evaluates both discrete and dimensional emotional models. Section 4 reviews information integration strategies in MER. Section 5 deals with the studies conducted in MER, and in part 6, MER DBs are introduced. Evaluation criteria and accuracy of models are presented in Section 7, and the importance of ethics in identifying emotions is mentioned in Section 8. Section 9 summarizes this article, while Section 10 provides final tips and suggestions. A table of abbreviations is presented in Table 1 to increase brevity and clarity.

1.1 | Review Methodology

This section offers an in-depth overview of the academic research conducted in the field of ER, providing a systematic review of the latest findings and delineating current research areas.

1.2 | Search Methodology

To orchestrate the Systematic Literature Review (SLR), a set of search queries was carefully selected, encompassing 'ER', 'MER', as well as 'ML' and 'DL'. The exploratory phase commenced with an automated query on Google Scholar for pertinent conference and journal publications. This phase was succeeded by meticulous manual searches through specialized DBs, engaging composite search strings such as "ER" conjoined with either "DL" or "ML" and similarly structured queries specifically targeting 'MER'.

1.3 | Criteria for Inclusion and Exclusion

Specific criteria were meticulously applied to ascertain the relevance of primary research works in the SLR. The data considered were sourced from scientific texts written in English and published in digital platforms such as IEEE, ACM, ScienceDirect, and Springer between 2013 and 2023, predicated on inclusion and exclusion criteria. Inclusion criteria encompassed the availability of the complete document for download, English language text, publication between 2013 and 2023, relevance to the research questions, coverage of ER methods, particularly focusing on multimodal emotion detection using DL, and accessibility via high-resolution digital DBs. We used the following keywords: DL, ER, text, visual, speech, audio, multimodal, trimodal, and bimodal. Exclusion criteria involved the unavailability of the complete document, documents in a language other than English, and publications outside the specified time frame.

1.4 | Selection of Studies

Following the prescribed search strategy and the inclusion and exclusion criteria application, around 640 research papers were identified as relevant, with 230 publications and conference articles being prioritized following a comprehensive review.

1.5 | Research Inquiries

This research aims to investigate the following inquiries:

I. What are the methodologies for enhancing recognition accuracy in emotion classification?

II. How can different emotions be analyzed and classified effectively?

III. What techniques in data fusion contribute to enhanced accuracy in recognizing emotions?

IV. What tools and technologies are suitable for acquiring data in ER?

V. What are the existing ER datasets and their key characteristics?

2 | Related Works

This section analyzes scholarly articles on affective computing published between 2013 and 2023, focusing on non-contact methodologies. *Table 2* offers a comprehensive summary of these recent reviews. The articles are compared based on sentiment model, DB, multivariate combination method, and Quantitative Evaluation (QE) criteria. As evidenced in *Table 2*, the various facets of affective computing are systematically delineated.

2.1 | Investigations Pertaining to Non-Contact Physical Sensing Methods

In studies pertaining to ER without direct physical interactions, the predominant approaches encompass visual, textual, and auditory techniques [12]. Within the visual domain, a significant portion of the literature is dedicated to FER [38], [39], [41], [44], Facial Micro-Expression Analysis (FMEA) [40], and 3D FER [41]. Poria et al. [45] further delved into discerning emotional representation from body gestures and subsequently embarked on a multimodal approach to ER, integrating both speech and combined facial and body movements. Within the textual realm, factoring in the underlying emotions embedded in dialogue can augment sentiment analysis and emotion detection efforts [42], [43], [46]–[48]. To address inherent challenges in emotion-centric AI platforms. Bota et al. [49] introduced a DL-based adversarial training paradigm leveraging three distinct physical signals. Conversely, Rouast et al. [12] examined the potential efficacy gains achieved by detecting the influence of multimodality across various fusion paradigms, juxtaposing them against singular modality analysis with a focus on multimodal fusion.

Table 2's reviews on emotion detection devoid of physical interaction predominantly highlight the use of DL techniques, underscoring its appropriateness for this domain. However, the current literature does not fully encapsulate the cutting-edge advancements and scholarly accomplishments in DL-driven emotional computing.

Table 1. List of abbreviations used in this article.

Acronym	Full Form	Acronym	Full Form
TSA	Textual Sentiment Analysis	SER	Speech Emotion Recognition
FER	Facial Expression Recognition	FMER	Facial Micro-Expression Recognition
4D/3D FER	4D/3D Facial Expression Recognition	EBGR	Emotional Body Gesture Recognition
EEG	Electroencephalogram	ECG	Electrocardiography
EMG	Electromyography	EDA	Electro-Dermal Activity
ML	Machine Learning	DL	Deep Learning
GMM	Gaussian Mixture Model	MLP	Multi-Layer Perceptron
NB	Naive Bayesian	LSTM	Long-Short-Term Memory
LDA	Linear Discriminant Analysis	DCNN	Deep Convolutional Neural Network
DT	Decision Tree	CNN	Convolutional Neural Network
KNN	K-Nearest Neighbors	RNN	Recurrent Neural Network
HMM	Hidden Markov Model	GRU	Gated Recurrent Unit
ANN	Artificial Neural Network	AE	Auto-Encoder
PCA	Principal Component Analysis	GAN	Generative Adversarial Network
MLP	Multi-layer Perceptron	VGG	Visual Geometry Group
SVM	Support Vector Machine	DBN	Deep Belief Network
RBM	Restricted Boltzmann Machine	HAN	Hierarchical Attention Network
RBF	Radial Basis Function	ResNet	Residual Networks
FC	Full-Connected	GAP	Global Average Pooling
MKL	Multiple Kernel Learning	AUs	Action Units
RF	Random Forest	AAM	Active Appearance Model
ICA	Independent Component Analysis	LFPC	Logarithmic Frequency Power Coefficient
BoW	Bag-of-Words	ROIs	Regions of Interest
LBP-TOP	Local Binary Pattern from Three Orthogonal Planes	MFCC	MEL Frequency Cepstrum Coefficient

2.2 | An Examination of Emotion Detection through Physiological Signal Analysis

In the realm of PER, advances have been facilitated by incorporating embedded devices that capture physiological signals, a process termed physiological coupling. García-Martínez et al. [50] provided an exhaustive review of the utilization of ML in ER by examining various physiological signals, offering insights into core principles, historical context, methodologies, and forthcoming advancements. Ekman [51] delved into the realm of nonlinear EEG-based ER and pinpointed specific nonlinear markers for prospective scholarly inquiries. Kim and André [29] undertook a systematic review of EEG-based ER literature spanning from 2009 to 2016, paying particular attention to aspects such as feature representation, classification techniques, and performance metrics. As delineated in *Table 2*, every review centered on physiological-based ER scrutinizes ML methodologies in discrete and dimensional emotional frameworks. Nonetheless, a single study has explored various DL strategies. This present review encapsulates the pivotal contributions of both ML and DL paradigms to the evolution of PER in the academic landscape.

Table 2. An overview of reviews related to emotional computing.

Author	Emotion Model		DB	Modality			Multimodal Fusion			Method		
	Dis	Dim		T	A	V	VA	TA	VAT	ML	DL	QE
Reviews on Physical-Based Affect Recognition												
[12]	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
[38]	×	×	✓	×	×	✓	×	×	×	✓	✓	✓
[40]	×	×	✓	×	×	✓	×	×	×	✓	✓	✓
[45]	✓	✓	✓	×	✓	✓	✓	×	×	✓	✓	×
[49]	×	×	×	✓	✓	✓	×	×	×	×	✓	×
[42]	×	×	✓	✓	×	×	×	×	×	✓	✓	✓
[46]	✓	✓	✓	✓	×	×	×	×	×	×	✓	✓
[47]	✓	✓	✓	✓	×	×	×	×	×	✓	✓	×

Table 2. Continued.

Author	Emotion Model		DB	Modality			Multimodal Fusion			Method		QE
	Dis	Dim		T	A	V	VA	TA	VAT	ML	DL	
Reviews on Physical-Based Affect Recognition												
[43]	×	×	✓	✓	×	×	×	×	×	×	✓	✓
[48]	✓	✓	✓	✓	×	×	×	×	×	✓	×	×
[41]	×	×	✓	×	×	✓	×	×	×	✓	✓	✓
Reviews on Physical-Physiological Fusion for Affective Analysis												
[13]	✓	✓	✓	×	✓	✓	✓	×	×	✓	✓	✓
[20]	✓	✓	✓	✓	✓	✓	✓	×	×	✓	✓	✓
[17]	×	×	✓	✓	✓	✓	✓	×	✓	✓	✓	✓
[14]	✓	✓	✓	✓	✓	✓	✓	×	✓	✓	✓	✓
[52]	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

*Multimodal Fusion: VAT=Visual-Audio-Textual, VA=Visual-Audio, TA=Textual-Audio, P-P=Physical-Physiological.
Modality: V=Visual (Facial expression, Body gesture), A= Audio (Speech), T=Textual

2.3 | Exploring Fusion Methods without Physical-Physiological Connections for Emotion Analysis

Recently, there has been a discernible inclination towards employing signals devoid of direct physical-physiological links to analyze emotions. Shoumy et al. [13] conducted a comprehensive review of DL-oriented methods for human ER that harness audio, visual, and physiological signals to extract spatial, temporal, and joint feature representations. Subsequently, Poria and Cambria [20] shed light on various methodologies pertaining to feature extraction, dimensionality reduction, and the employment of ML-driven classifiers, particularly within the realm of multi-channel EEG ER.

This study further delineated recent advancements in MER utilizing ML or DL techniques. Corneanu et al. [17] offered insights into multimodal datasets, emphasizing feature extraction techniques based on EEG, imagery, audio, and textual information while also exploring multimodal fusion tactics and real-time ER procedures. Ekman [14] systematically reviewed diverse frameworks and methodologies that utilize textual, audio, visual, and physiological signals, providing a thorough performance evaluation. Their work culminated in a discussion on the myriad applications within sentiment analysis, subsequently highlighting emerging trends and prospective avenues of research. The seminal contributions in *Table 2* delve into the core tenets of sentiment computation, encompassing sentiment modeling and both unimodal and multimodal detection approaches via ML and DL paradigms. However, a critical appraisal of the existing literature reveals a void regarding comprehensive comparative analyses within unimodal and multimodal sentiment analysis domains.

3 | Emotional Models

In the realm of emotional computing, it is imperative to define and understand the concepts of emotions or feelings. The foundational notion of emotion was pioneered by Ekman [51], [53]. While experts from diverse disciplines, including neuroscience, philosophy, and computer science, have endeavored to categorize emotions through different lenses [54], a universally agreed-upon model of emotion remains elusive. Notwithstanding, two predominant emotion models are commonly employed in emotional computing: the discrete emotion model [53] and the continuous emotion model [16], [55].

3.1 | Discrete Emotion Model

In the domain of emotional study, the discrete emotion paradigm, often called the categorical emotion paradigm, classifies emotions into distinct categories. Among the most prominent discrete emotion frameworks are Ekman's Six Basic Emotions [53] and Plutchik's Emotion Wheel model [56], as depicted in *Fig. 3(a)* and *Fig. 3(b),* respectively. The foundational model by Ekman [51] and its subsequent derivatives [57],

[58] have gained considerable traction within the ER scholarly community [59], [60]. These six fundamental emotions, typically encompassing anger, disgust, fear, happiness, sadness, and surprise, are conceptualized based on several criteria [53]:

I. Such emotions are rooted in human instinct.

II. Given identical circumstances, individuals tend to manifest analogous basic emotions.

III. It's posited that people convey these core emotions with congruent semantic interpretations.

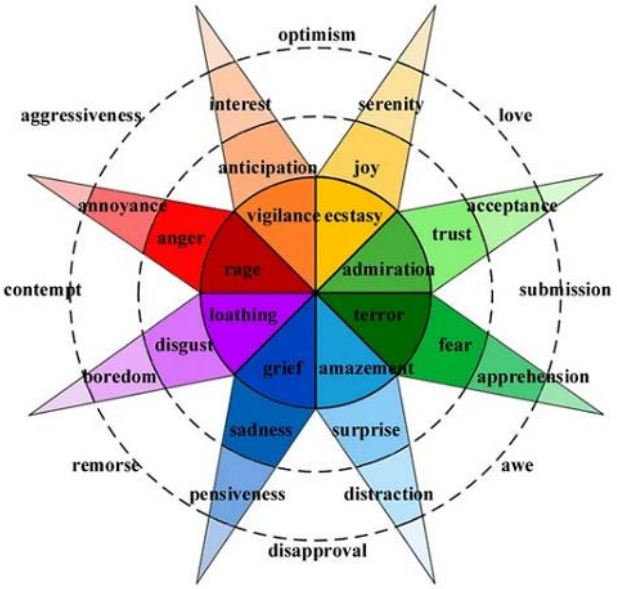
IV. There's a contention that these quintessential emotions manifest in a consistent expression pattern across diverse individuals.

Ekman's foundational model underscores the universality of human emotions across ethnicities and cultural spectra. However, nuances might arise in perceiving these basic emotions contingent on cultural contexts, and the amalgamation of various basic emotions can give rise to intricate or composite emotional states [15]. Conversely, Plutchik's wheel-based paradigm [56] encompasses eight foundational emotions: happiness, trust, fear, surprise, sadness, anticipation, anger, and disgust. This model elucidates the interconnections among these emotions, as visualized in *Fig. 3(b)*. To illustrate, happiness and sadness are posited as antithetical emotions, while anticipatory sentiments could metamorphose into states of alertness. This concentric representation, frequently denoted as the component model, predicates that central emotions exhibit heightened intensity, whereas peripheral emotions reflect their respective intensity gradient. Such discrete emotions are typically segmented into three overarching polarities: positive, negative, and neutral, which are instrumental in sentiment analysis. To enhance the precision in emotion delineation, discrete emotion management [61] has been proposed, emphasizing multilayered emotion analysis, subsequently augmenting classification efficacy.

3.2 | The Lateral Emotions Model

Numerous scholars have gravitated toward a continuous multidimensional framework in addressing the constraints of discrete emotion paradigms. The Pleasure, Arousal, and Dominance (PAD) model [16], illustrated in *Fig. 4,* serves as a salient example. This model mirrors Mehrabian's tripartite spatial theory of emotions [62] by encapsulating three distinct dimensions:

a.



b.

Fig. 3. Two distinct models for emotion computing: a six fundamental emotional models [15], illustrated through emojis; b) partial models, such as Palachik's emotion wheel model [56].

The Pleasure dimension (often referred to as Valence) spans the emotional spectrum from profound distress to euphoria. The Arousal (or Activation) dimension quantifies the extent of physiological activity and cognitive alertness. The Mastery (or Attention) dimension conveys the sentiment of either being swayed by external stimuli, including individuals or having an impact upon them.

Dabas et al. [63] posit that the PAD model's pleasure and arousal dimensions can encapsulate the vast emotional nuances. Consequently, Plutchik [55] introduced a circular framework hinging on arousal dynamics to depict intricate emotional states. This design delineates a continuous, bidimensional emotional realm demarcated by axes of Valence (ranging from pleasantness to its antithesis). The circular construct is divided into four sectors. The inaugural sector, signifying positively-valenced activation, aligns with emotions akin to joy. Contrastingly, the third sector, marked by subdued arousal and detrimental sentiments, aligns with despondency. The penultimate quadrant, epitomizing heightened arousal juxtaposed with negative sentiments, resonates with anger, while the concluding quadrant, characterized by minimal arousal and positive sentiment, embodies tranquillity [64]. This figure presents the Pleasure-Arousal-Dominance (PAD) model [62].

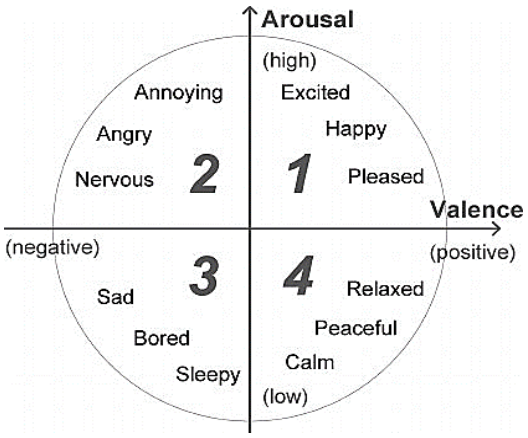


Fig. 4. Subsequent models of emotions.

4 | Multimodal Information Fusion Strategies for MER

Domains and modalities are crucial in the field of ER. Although each domain and modality has unique features, similarities and differences exist. All domains and modalities require data collection, processing, feature extraction, and classification. Additionally, they share the common objective of recognizing and interpreting human emotions. However, it is important to note differences in data collection methods, feature extraction techniques, and classification algorithms across domains and modalities. Each domain and modality presents its own unique set of challenges and limitations that researchers must address. Therefore, it is crucial to clearly understand the similarities and differences among domains and modalities to develop effective ER systems. Emotions are complex phenomena that involve various physiological, behavioural, and cognitive processes. ER systems that rely on a single modality or domain may not be sufficient to capture the complexity of emotions. MER, which combines information from multiple modalities or domains, has improved the accuracy and robustness of ER systems.

Additionally, MER can provide a more comprehensive understanding of emotions by capturing different aspects of emotional expression. For instance, facial expressions can convey the valence and intensity of an emotion, while speech can provide information about the content and context of an emotion. Moreover, combining modalities can help overcome the limitations and challenges of individual modalities, such as noise, variability, and ambiguity. Therefore, MER has become an active and promising research area in affective computing. This section provides an overview of MER, including the primary fusion techniques used in this area of research, as well as a review of recent studies. Multimodal learning can change various modal information by jointly modeling various modal data and identifying their internal association. Creating a shared representation space is critical because associating diverse modes is the fundamental challenge of multimodal learning. Fusion and alignment are two basic techniques for creating a common representation space [65].

4.1 | Information Fusion

Multimodal fusion combines information from multiple modalities to enhance its accuracy. Multimodal alignment is used to find the relationship between different modalities of information. This technology is employed in the study of ER to combine data such as audio and video, which increases recognition accuracy. Zhang et al. [65] identified three fusion methods in ER: feature-level, decision-level, and model-level.

The most commonly used method in MER, according to [66], is feature-level fusion, which involves combining the features extracted from each modality into a new feature vector. The feature vector for emotion identification is often of a larger dimension, which is then reduced using dimension reduction methods. A classifier is then used to identify the emotion. To fuse the feature layer, the relationships between various modalities are exploited. However, this method does not consider the distinction between the emotional characteristics of different modalities. Achieving time synchronization between multiple modalities is also challenging. Learning the relevance across modalities has become increasingly challenging due to the diversity of modalities [11].

The decision-level fusion method treats each modality independently before combining the results [65]. This involves classifying features from different modalities using their respective classifiers. The final result's recognition accuracy is improved by combining the results of different UERs using an algebraic combination rule [66]. Poria et al. [11] emphasize the difference between various features in decision-level fusion and suggest applying the most suitable classifier for each modality. However, this approach does not consider the relevance between features, and the learning process is time-consuming.

Model-level fusion is another option. Jiang et al. [16] reviewed fusion methods in multimodal recognition and found that model-level fusion is less commonly used than feature-level and decision-level fusion methods. Model-level fusion relies on data correlation and a relaxed data fusion approach across several modalities. The realization of model-level fusion depends on the models used. The results of recognition are obtained

through a multimodal information-combining model that learns the multimodal interaction within the model and builds a shared space of representation.

4.2 | Multimodal Fusion Strategies for MER

The fusion of diverse modalities holds pivotal importance in MER. Common multimodal information fusion strategies encompass feature-level fusion, decision-level fusion, and model-level fusion. Notably, model-level fusion has gained prominence due to its superiority over the other methods. Cross-modal attention-based fusion strategies have recently garnered attention for effectively capturing inter-correlation across different modalities. Feature-level fusion, illustrated by concatenation, represents the simplest fusion method but exhibits limitations in capturing inter-modality associations.

Conversely, decision-level fusion endeavors to model each input modality independently and integrate the outcomes using algebraic rules. However, it often fails to capture the inter-modality relationships. Model-level fusion consistently demonstrates superior performance among these methods, highlighting the need for advanced model-level fusion strategies in MER. Notably, many existing MER approaches have not fully exploited temporal and semantic alignment information across different modalities. The effective integration of temporal and semantic alignment for reconciling emotional-related information remains an underserved area. Hence, there is a critical need for further exploration into mechanisms for capturing temporal and semantic alignment across different modalities in various MER systems. In conclusion, addressing the aforementioned research challenges in lightweight and explainable deep models and developing advanced multimodal information fusion strategies is indispensable for progressing multifaceted ER.

4.3 | Cross-Corpus MER

Despite significant advancements in single-corpus MER studies, challenges persist in the cross-corpus setting, where training and testing corpora originate from different sources. Existing MER models encounter difficulties in generalizing to new corpora due to corpus-specific biases and substantial variations in feature distributions. These discrepancies arise from differences in recording device quality, spoken languages, cultural nuances, and emotion annotation, among other factors. Consequently, applying existing MER models to new corpora results in significant performance degradation. Current efforts have predominantly focused on cross-corpus single-modal ER, particularly in audio and FER. However, limited research has specifically addressed cross-corpus MER, with particular attention to mitigating cross-cultural discrepancies. Notable experiments utilizing adversarial learning frameworks, as observed in the work by Liang et al. [64], present a promising approach to addressing the cultural influence on MER tasks. Future research endeavors may delve deeper into applying adversarial learning methods for cross-corpus MER, particularly those based on Generative Adversarial Networks (GANs).

4.4 | Additional Modalities for MER

The prevalent focus of earlier MER research on audio, visual, and textual modalities may not fully capture the comprehensive features required for comprehensive emotional state characterization. Consequently, integrating additional modalities is a compelling direction for advancing MER. Beyond physiological signals, body language, encompassing gestures, postures, and eye movement, represents a noteworthy visual modality for expressing human emotions. Although challenges related to gender differences, cultural influences, and scant-labeled datasets exist, the construction of multimodal emotional datasets integrating body language with other modalities holds significant promise for future research in MER. Consequently, investigating the performance of MER by integrating body language with other modalities using these constructed datasets warrants dedicated exploration.

4.5 | Few-Shot Learning for MER

Conventional MER studies extensively rely on a large volume of annotated multimodal emotional samples, particularly when employing DL for feature extraction. However, procuring extensive annotated multimodal

emotional samples poses significant challenges regarding manpower and resources. Furthermore, advancements in psychology have led to the development of new conceptual and computational methods for representing the intricate "semantic space" of emotion, resulting in a nuanced taxonomy comprising 27 distinct emotional categories. Consequently, the evolving diversity and granularity of emotion categories have significantly complicated the collection of annotated emotional samples. To tackle these challenges, there is a burgeoning interest in exploring effective few-shot learning strategies for MER. Recent endeavors have delved into investigating the few-shot learning scenario for MER. Notably, Caridakis et al. [67] proposed a prompt-based multimodal fine-tuning method for few-shot MER tasks, emphasizing a unified pre-training strategy with two stages to bridge the semantic gap between text and visual modalities. This underscores the importance of developing appropriate multimodal pre-training strategies for few-shot MER, reflecting a compelling area for future research. Furthermore, exploring advanced zero-shot learning approaches for MER, targeting the identification of rare unseen emotions, represents a crucial research direction.

4.6 | Attention Neural Networks for MER

Attention neural networks [68] are DL architectures equipped with an attention mechanism. This mechanism allows the network to focus on specific parts of input data, improving its performance in various tasks. The first attention neural network was the attention-based RNN [69]. Recently, there has been extensive interest in transformer techniques that leverage a unique self-attention mechanism [70] due to their strong ability to model long-term dependencies. Various transformer-based methods have been developed in recent years, including the vision transformer [71], audio [72], and video transformer [73]. Furthermore, transformer-based techniques have been effectively utilized in various applications, such as image classification [74], object detection and segmentation [75], speech signal processing [52], automated depression detection [76], and air quality prediction [77]. However, DL methods are rarely used for MER despite their ability to effectively learn high-level features from input data. However, DL methods are rarely used for MER despite their ability to effectively learn high-level features from input data. However, DL methods are rarely used for MER despite their ability to effectively learn high-level features from input data. It has been pointed out that transformer-related methods have not been widely adopted in this field. Although computational complexity can be an issue for many DL methods, they remain a promising approach for multimodal emotion classification.

5 | Overview of MER Studies

Recent studies have indicated the superiority of MER over single-modal approaches. Incorporating audio, visual, text, and other modalities is essential for achieving accurate MER. Traditional multimodal fusion methods encompass feature-level fusion (early fusion), decision-level fusion (late fusion), model- and hybrid-level fusion. Early fusion directly consolidates features from individual modalities into a singular feature vector for emotion classification. Still, it may encounter issues related to dimensionality and fail to consider temporal associations across modalities. In contrast, late fusion combines outcomes from different modalities using mathematical rules such as "majority vote" or "average," enabling each modality to employ its classifier. However, this approach treats modalities as independent entities and does not capture their interconnectedness. Model-level fusion individually models each modality while accounting for intermodal correlations, facilitating interconnectedness and minimizing the necessity for modal time synchronization. Additionally, hybrid-level fusion combines diverse fusion strategies, capitalizing on the strengths of different fusion approaches. Common MER techniques are classified into bimodal and trimodal ER using single-modal information. These fusion methods are assessed within the context of bimodal and trimodal ER, and the findings are outlined in *Table 3*. In this section, we elucidate the amalgamation of multimodal signals to establish a framework dedicated to multimodal sentiment analysis [78]. It is imperative to consider various methodologies [79]. The synthesis of multimodal analysis can amplify accuracy and proffer a more holistic comprehension in contrast to unimodal sentiment identification [80], [81]. Noteworthy reviews centered on multimodal sentiment analysis [12], [14], [17] predominantly emphasize strategies for multimodal fusion. Such strategies can be demarcated into feature-level fusion (early fusion), decision-level fusion (late fusion), model-

and hybrid-level fusion. Yet, it's worth noting that multimodal sentiment analysis might also employ various techniques. Multimodal emotion analysis can be categorized into fusion strategies for emotion analysis from multimodal sources, fusion of multiphysiological techniques dedicated to emotion analysis, and fusion integrating physical and physiological models for emotion analysis. The classification of these categories is grounded on four distinct fusion strategies. Salient exemplars of these fusion strategies are illustrated in *Fig. 5*.

Feature-level fusion

Feature-level fusion is a technique that amalgamates features derived from diverse input sources to formulate a comprehensive feature vector. This aggregated vector is subsequently introduced to a classifier for further processing. Illustrations of this type of fusion, particularly for models incorporating combinations of video, audio, and text inputs, are presented in *Fig. 5(a)* to *Fig. 5(c)*.

Model-level fusion

Model-level fusion is an approach that identifies correlation patterns amongst features extracted via different methodologies. This method necessitates the utilization of a fusion model, exemplified by Hidden Markov Models (HMM) or Extreme Learning Machines (ELM) Convolutional MKL-based multimodal emotion recognition and sentiment analysis [82]. For clarity, *Figures 5(d) and 5(e) depict model-level integration applied to physical-physiological* and video-audio-text models, respectively.

Hybrid fusion

Hybrid fusion denotes the synthesis of both feature-level and decision-level fusion techniques. It effectively merges the advantages of these two methodologies. A representative example showcasing the hybrid amalgamation of video, audio, and text modalities can be viewed in *Fig. 5(g)* and *Fig. 5(d)* [69]–[71], [83].

5.1|Employing a Contact-Free Multimodal Approach for Sentiment Analysis

In the realm of sentiment analysis, a prevalent approach entails the amalgamation of disparate unimodal techniques. These can be delineated as visual-audio sentiment recognition [31], [84], text-audio sentiment recognition [82], [85], and a tri-modal approach encompassing visual, audio, and text sentiment recognition [82], [84]. An overview of conventional strategies for multi-modal emotion analysis can be found in *Table 3*, which will be elaborated upon in the subsequent sections.

5.1.1|Audio-visual emotion recognition

In day-to-day interactions, emotions are primarily conveyed through visual and auditory cues [86]. Studies [81], [87]–[90] indicate that the combined recognition of emotions using both visual and auditory indicators is superior to their individual use. An overview of audio-VER is given in *Table 3*. The architecture of the input of audio and video information, the use of the essential audio and video feature extraction methods, and the type of feature fusion are shown in *Fig. 6*.

Feature-level fusion

Zhao et al. [91] introduced an ML approach for audio-VER, integrating HOG-TOP, acoustic, and geometric attributes, as depicted in *Fig. 5(a)*. These features are subsequently processed through a multi-core SVM for improvisational and predetermined FER. Meanwhile, Baltrusaitis et al. [31] employed a Convolutional Neural Network (CNN) and a deep network to extract these cues, which were further processed by a two-layer Long Short-Term Memory (LSTM) model to estimate Arousal-Valence values. Noteworthy techniques for identifying audio-visual emotions have been documented in [92] and [93].

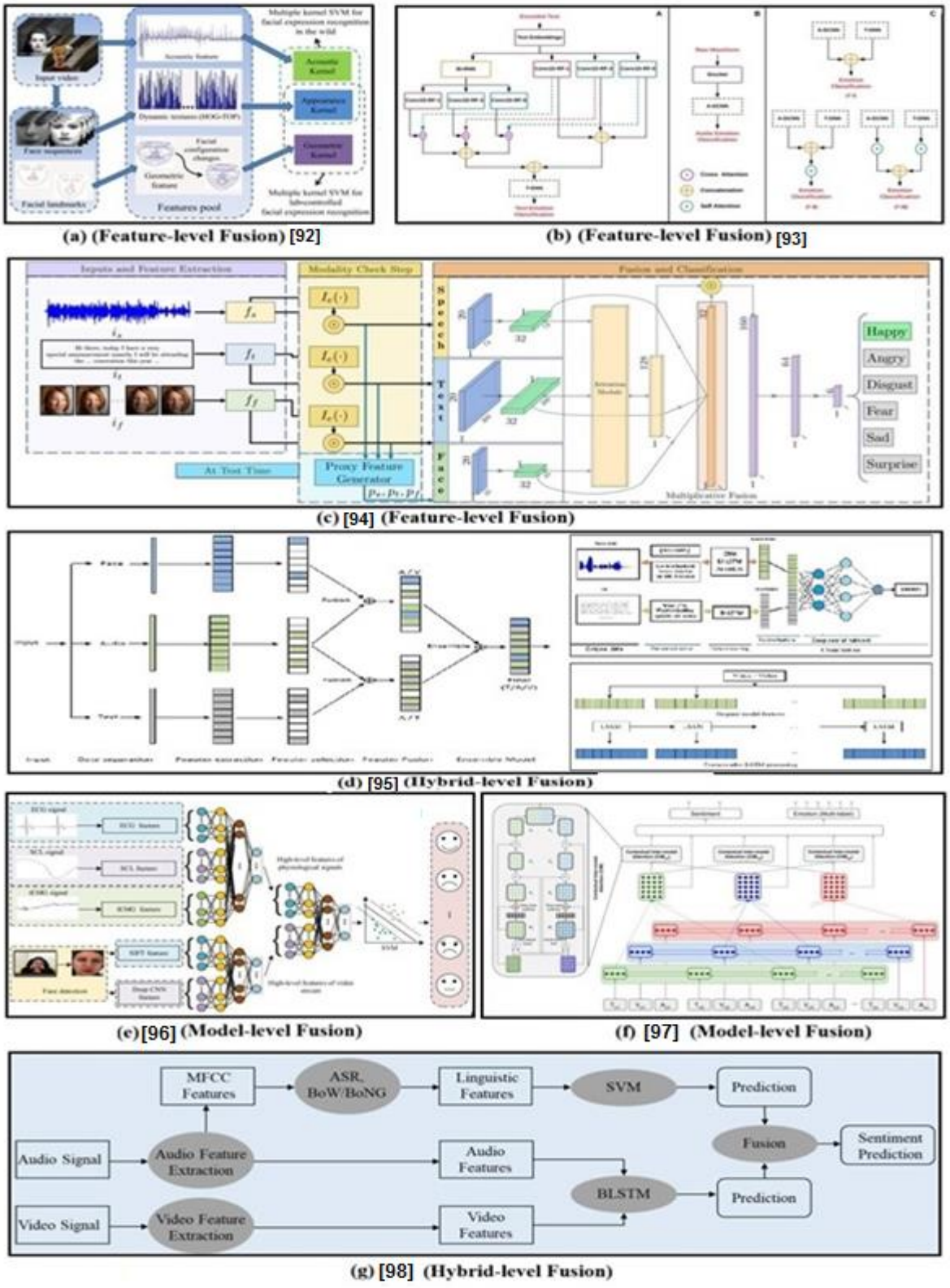


Fig 5. The classification of MER is described in the reference.

Fig. 5(a): integration of features is used for visual-audio emotion recognition, described further in reference [92]. *Fig. 5(b)*: integration of features is used for audio-text emotion recognition, described further in reference [93]. *Figs. 5(a)* and *5(d)*: feature-level integration recognizes emotions from video, audio, and text [94], [95]. *Fig. 5(e)*: model-level fusion is used for analyzing physical and physiological emotions [96]. *Fig. 5(f)*: model-level integration recognizes emotions from video, audio, and text [97]. *Fig. 5(g)*: combined-level integration recognizes emotions from visual, audio, and text sources [98].

In particular, Hara et al. [93] introduced a method to determine emotion-centric regions using various techniques. They recommended a Factorized Bilinear Pooling (FBP) fusion model to fuse video-audio attributes, compensating for disparities in video frame expressions. This technique resulted in a recognition precision of 62.48% in the EmotiW2018 AFEW audio and video segment. It pioneered a deep Visual-Audio Attention Network (VAANet), emphasizing unique attention components and polarity-adaptive entropy loss. Remarkably, the spatial, channel, and temporal attentions were amalgamated with a 3D CNN [97] for spatial and video sequence parsing.

Decision-level fusion

Ramzani Shahrestani et al. [98] presented a visual-audio ER structure that synergizes multitasking with multifaceted feature learning. Specifically, they developed four sub-models using SVM and CNN classifiers for DL-oriented audio and visual cues. These models were later synchronized to infer the conclusive emotion through a hybrid approach. Carr achieved commendable accuracies of 81.36% (speaker-independent) and 78.42% (speaker-dependent) in the eNTERFACE [36] evaluation.

Model-level fusion

This fusion demands ML models, such as HMM, Kalman filters, and Deep Belief Networks (DBNs), to discern associations from varied techniques for ultimate determinations. Huang et al. [33] devised an HMM (SC-HMM) to correlate the temporal dynamics of audio and video signals, followed by a Bayesian classification. The SC-HMM with Bayesian classification reported accuracies of 90.59% (across four emotion categories) and 78.13% on the human-machine multimedia communication DB and the SEMAINE DB, respectively. Zhang et al. [36] conceptualized Kalman filters rooted in Markov's model, fusing temporally ordered classifier decisions for emotion identification. In this context, audio and video data are merged through unimodal feature extractors and foundational classifiers based on Kalman filters and trustworthy metrics. Ko [37] harnessed CNNs to distill audio and video cues, formulating a profound fusion strategy (DBNs) to integrate features. This was coupled with a linear SVM classifier, registering accuracies of 80.36%, 54.57%, and 85.97% across RML, eNTERFACE05, and BAUM-1 datasets.

Similarly, Zhang et al. [36] proposed a mixed score-level method to compute all likelihoods of DBNs trained on the spatio-temporal dynamics of audio-visual channels. Priyasad et al. [94] utilized two-dimensional and three-dimensional CNNs to extract advanced features from refined audio and video cues appropriate layers. Similarly, Zhang et al. [36] proposed adopting a conversational transformer model to capture intra-modal and cross-modal interactions, employing a conversational transformer, Bi-GRU network, speaker embedding, and ATS-Fusion component for audio-text fusion, ultimately facilitating emotion classification. Krishna et al. [99] presented an Adaptive Interactive Attention Network (AIA-Net) for audio-text ER, focusing on the dynamic interactive interrelations between text and speech feature representations and their subsequent utilization for emotion classification extraction. This was complemented by utilizing a cross-attention mechanism for N-gram level correlation learning and a self-attention-based fusion method for ER from the concatenated audio and text features, culminating in integrating FC and softmax layers. Hazarika et al. [100] innovatively introduced a novel MER approach founded on cross-modal attention and 1D-CNNs, employing audio and text encoders to obtain high-level feature representations and a cross-modal attention network for interlacing audio and text sequences, leading to emotion classification using.

Table 3. An overview audio-VER.

Publication	Feature Representation	Classifier	Fusion Strategy	Database	Performance
[31]	CNN, Resnet	LSTM	Feature-level	RECOLA	A/V: 78.80/73.20
[98]	C3D + DBN	Score-level fusion	Model-based	eNTERFACE	6 classes: 89.39
[91]	Acoustic, Geometric and HOG-TOP	Multiple kernel SVM	Feature-level	CK; AFEW	7 classes: 95.7 7 classes: 45.20
[84]	2D CNN, 3D CNN	ELM-based fusion, SVM	Model-based	Big Data eNTERFACE	3 classes: 91.30 6 classes: 78.42
[98]	Multitask CNN	Meta-Classifier	Decision-level	eNTERFACE	6 classes: 81.36
[92]	2D ResNet+Attention 3D esNet+Attention	FC	Feature-level	VideoEmotion-8 Ekman-6	8 classes: 54.50 6 classes: 55.30
[36]	3D CNN	Soar	Feature-level	eNTERFACE	6 classes: 88.1
[101]	Audio: CNN Visual: 3D-CNN	Linear SVM	Model-level	1)RML 2) eNTERFACE'05 3)BAUM-1 s	1)Acc(6-class):80.36% 2)Acc(6-class):85.97% 3)Acc(6-class):54.57%
[102]	Audio: LLDs Visual: Geometric Features	Linear FC	Model-level	AVEC-2017	CCC(Arousal):0.654 CCC(Valence):0.708
[103]	Audio: CNN Visual: VGG16	FC + Softmax	Model-level	eNTERFACE'05	1)Acc(6-class):80.83% 2)F1-score:80.23%
[104]	Audio: FCN Visual: VGG-face	FC + Softmax	Model-level	1)IEMOCAP 2)AFEW8.0	1)Acc(4-class):75.49% 2)Acc(7-class):63.09%
[94]	Audio: CNN Visual: CNN	FC + Softmax	Hybrid-level	1)SAVEE 2)RAVDESS	1)Acc(7-class):99% 2)Acc(8-class):86%
[81]	Audio: LLDs Visual:CNN + Bi-LSTM	FC + Softmax	Feature- level	1)SAVEE 2)RAVDESS 3)RML	1)Acc(7-class):99.75% 2)Acc(7-class):94.99% 3)Acc(7-class):99.23%

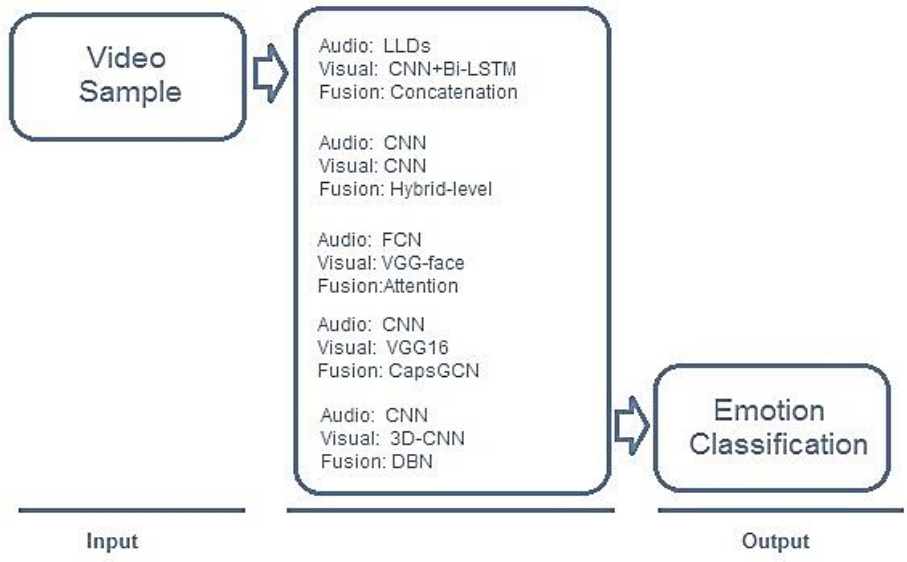


Fig. 6. Depicts an overview of the framework for diverse audio-VER, encompassing audio-visual feature extraction and fusion methods.

5.1.2 | Recognising textual and audio emotions

Recent advancements in audio-text ER have led to the development of several prominent algorithms, including CNN, RNN, LSTM, GRU, AIA-Net, and GCN. For instance, Priyasad et al. [94] proposed a feature-level fusion technique that integrates self-attention for audio-text ER, entailing the extraction of high-dimensional hand-crafted LLDs for speech signals and textual feature derivation utilizing CNNs. Fusing these features using a weighted addition based on attention score values and the subsequent employment of FC and softmax layers facilitated effective emotion classification. Fu et al. [105] also introduced a DL-based approach to fuse audio and text modalities for ER, leveraging the SincNet layer for speech feature extraction and a combination of CNN and Bi-RNN followed by a CNN for text feature. Moreover, Yoon et al. [106] developed an audio-text ER method based on context and knowledge-aware Graph Convolutional Networks (GCN), involving the deep speech feature extraction, fine-tuning of pre-trained BERT models for textual feature extraction, and the integration of an emotional lexicon into building knowledge graphs, ultimately leading to effective ER. By adopting these cutting-edge methodologies, researchers have significantly advanced the field of audio-text ER, introducing innovative techniques that hold promise for real-world applications.

Feature-level fusion

Cai et al. [107] introduced a dual deep recurrent neural network to encode sequences integrating both text and audio information. These outputs were then merged to deduce the final sentiment. Using the IEMOCAP dataset, this model attained an accuracy rate of 71.8% across four categories. Wu and Liang [108] developed an enhanced CNN combined with a Bidirectional Long-Short Term Memory (Bi-LSTM) to extract spatial attributes and capture their temporal dynamics, utilized band-pass filters in conjunction with a Deep Convolutional Neural Network (DCNN) to extract both textual and auditory characteristics. They integrated varied methodologies to predict four emotions in the IEMOCAP dataset using text-acoustic features.

Decision-level fusion

Jin et al. [109] incorporated a Meta-Decision Tree (MDT) to integrate acoustic-prosodic data with classifiers like GMM, SVM, and MLP. Additionally, they employed the Maximum Entropy (MaxEnt) model to elucidate the correlation between states and emotional conveyance. Semantic tag rules are amalgamated to detect both speech and text sentiments, achieved by employing a weighted fusion strategy rooted in AP and SL for final sentiment prediction.

Feature-level fusion vs. decision-level fusion

New lexical and auditory attributes were created to ascertain the relative efficiency of feature- and decision-level fusion in audio-text ER [110]. Subsequently, two composite strategies were applied, yielding 55.4% and 69.2% recognition accuracies for four classes in the IEMOCAP dataset. Hazarika et al. [100] employed multiple dual RNNs to encode text-to-audio sequences, particularly in emotional dialogues encompassing auditory and textual elements. In sum, the efficacy of both feature-level and decision-level fusion techniques has been analyzed through three distinct methodologies to gauge their performance.

Though SER [26] and TSA [111] have made notable advancements utilizing distinct techniques, achieving compelling results using individual methods remains challenging [91]. Strategies that extract emotions from text-to-audio leverage both speech characteristics and linguistic content to augment the efficiency of unimodal emotion detection systems [35], [111]. An overview of text-audio ER is given in *Table 4*.

In recent years, the research focus on Trimodal Emotion Recognition (TER) has led to the introduction of a diverse array of algorithms, such as MESM, MMAN, Bi-GRU, M3ER, CAT-LSTM, TransModality, MEmoBERT, IMAN, among others. These algorithms are built on comprehensive principles and approaches to address the complex task of MER. For instance, Dai et al. proposed a Multimodal End-to-end Sparse Model (MESM) designed for effective emotion classification, leveraging audio and visual. The architecture of

entering audio and video information and using the most important methods of audio and text feature extraction and the type of feature fusion is shown in *Fig. 7*.

Table 4. An overview text-audio ER.

Public	Feature Representation	Classifier	Fusion Strategy	Database	Performance
[109]	Acoustic-prosodic Semantic labels	Base classifiers, MDT, axEnt	Decision-level	2033 utterances	4 classes: 83.55 4 classes: 85.79
[91]	A-DCNN, T-DNN Self-attention	FC	Feature-level	IEMOCAP	4 classes: 80.51 4 classes: 79.22
[100]	Acoustic features Word embeddings	Pooling Scalar weight fusion	Decision- level Feature-level	IEMOCAP MSP-PODCAST	565.10/558.20 563.90/558.00
[94]	Audio: LLDs Text: CNN	FC + Softmax	Feature- level	IEMOCAP	Acc(4-class):71.4% F1-score:71.3%
[99]	Audio:SincNet+CNN Text: CNN + Bi-RNN	FC + Softmax	Model-level	IEMOCAP	Acc(4-class):80.51% WA:79.22% UA:80.51%
[111]	Audio: CNN + Bi-LSTM Text: Glove + CNN	FC + Softmax	Model-level	IEMOCAP	Acc(4-class):72.82%
[112]	Audio: LLDs Text: Lexical Features	FC + Softmax	Model-level	1)IEMOCAP 2)MELD	1) WAA(4-class):83.6% WAF1(4class):83.8% WAA(6-class):68.0% WAF1(6class):67.5% 2) WAA(7-class):62.0% WAF1(7class):60.5%
[105]	Audio:Wav-RoBERTa Text: RoBERTa	FC + Softmax	Model-level	1)IEMOCAP 2)MELD 3)CMU-MOSEI	1): Acc(4class):87.44% F1-score(4class): 87.16% 2) Acc(7-class):65.09% 3) 0.574 Acc(7class):53.20% Acc(2-class):89.33% F1-score:89.33%
[113]	Audio: CNN+Bi-LSTM Text: BERT	FC + Softmax	Model-level	1)IEMOCAP 2)MELD	1) Acc(4-class):85.82% , F1-score:85.90% 2) Acc(7-class):66.40% F1-score:64.63%

Features were extracted using a pre-trained VGG16 model and transformer for modeling temporal information. Additionally, a cross-modal sparse CNN block incorporating sparse CNN and cross-modal attention facilitated computational cost reduction, ultimately generating final emotion prediction scores through a weighted sum of classification scores from audio, visual, and text modalities. Subsequently, Pan et al. [114] introduced a Contextual Attention-based LSTM (CAT-LSTM) network designed to capture contextual information for utterances in multimodal sentiment analysis, highlighting the extraction of LLD features for the audio modality, utilization of a 3D-CNN for visual features, and pre-trained Word2vec for textual features. An attention-based fusion mechanism, culminating in an FC and Softmax layers, facilitated final emotion classification. Moreover, Mittal et al. [84] proposed a Multi-Modal Attention Network (MMAN)

for MER, utilizing typical LLD features, a 3D-CNN, and Word2vec for speech, visual, and text features, respectively. The MMAN integrated a multi-modal attention cLSTM for early fusion and three independent unimodal blocks for late fusion, complemented by an FC layer and a Softmax layer for emotion classification [115].

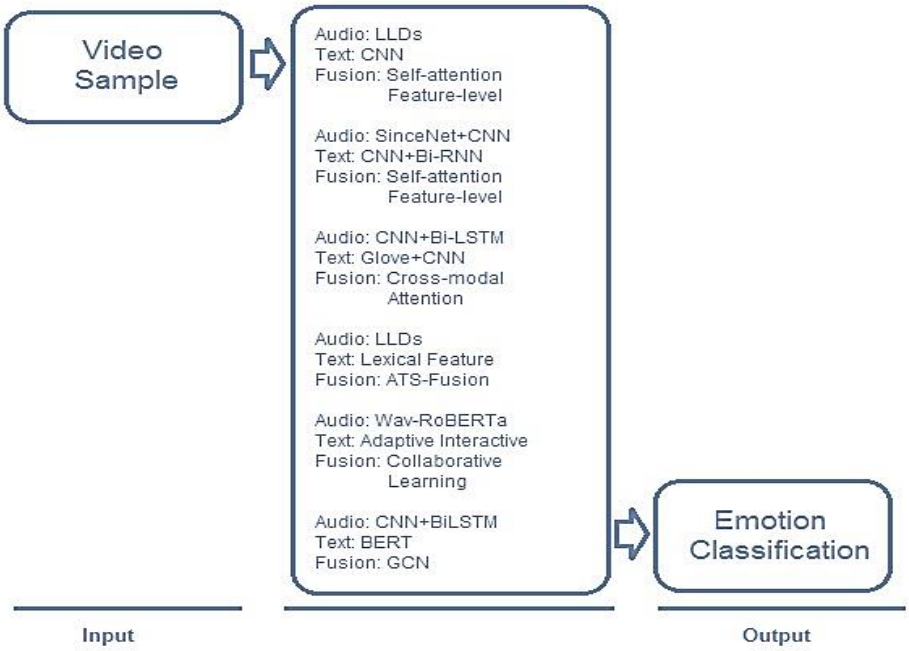


Fig. 7. Here is a summary of the framework for diverse audio-text ER involving audio-text feature extraction and fusion methods.

5.2 | Recognition of Emotions in Video, Audio, and Text

Developed a MER model, incorporating a multiplicative fusion layer to enhance emotion classification, focusing on popular LLD features for speech, facial action units and landmarks for visual features, and pre-trained Glove for text features. Furthermore, Metallinou et al. [116] introduced a multi-fusion residual memory network approach for identifying utterance-level emotions, leveraging a Bidirectional GRU (Bi-GRU) model and FC layers to learn temporal and intra-modality interaction cues, along with a multi-stage fusion module for emotion classification. Subsequent studies have introduced innovative approaches such as end-to-end fusion with Transformer (TransModality), Interactive Multimodal Attention Network (IMAN), multi-channel weight-sharing Autoencoder integrating cascade multi-head attention, hybrid contrastive learning for trimodal feature representations, and MEmoBERT, a pre-training deep model customized for MER, all of which demonstrate the robustness and complexity of trimodal ER research. Multimodal systems integrating voice, facial expressions, and text are pivotal in discerning genuine emotional states, as corroborated by various studies [117], [118]. For instance, a woman's response, "I do," accompanied by tears upon a proposal, can't be solely interpreted based on text, audio, or video. Thus, amalgamating these three inputs can potentially augment the accuracy of ER systems. An overview of video-audio-text ER is given in *Table 5*.

Feature-level fusion

Arguello and Rosé [119] utilized techniques like BoW, OpenEAR [120], and vision software for extracting linguistic, auditory, and visual cues, respectively. The trifold model encompassing text, auditory, and visual modalities showcased a marked advancement in efficiency, surpassing both singular and dual-modality models. Another study by Eskimez et al. [86] employed standard RNN, deep CNN, and openSMILE to

capture temporal image data, spatial text data, and audio data descriptors, respectively. Moreover, MKL was tailored to cherry-pick features from varied techniques to enhance detection. Lin et al. [85] introduced multiplicative M3ER. This method began by selecting feature vectors from three primary methods. Subsequent stages involved refining these features for optimization and eventually amalgamating them to predict six emotions. Within datasets like IEMOCAP and CMU-MOSEI, M3ER posted impressive accuracy rates of 82.7% and 89.0%, respectively. In contrast to the language-agnostic models in English or German sentiment analytics, Chinese sentiment analysis delves into explicit and implicit symbol meanings, e.g., phonetic orthography. Poria et al. [113] built on this by suggesting a DISA model, intertwining phonetic features with text and visual indicators.

Feature-level fusion vs. decision-level fusion

Savva et al. [121] employed CNN for textual extraction and openSMILE for auditory features, further synthesizing artificial image data features. As HOW [79] illustrated, feature and decision-level amalgamation use feature selection to detect uni and multimodal sentiments. Studies showed that while textual sentiment recognition reigns supreme among unimodal techniques, integrating visual-audio-text sentiment recognition trumps unimodal and bimodal techniques. However, feature-level fusion generally has superior accuracy, albeit at a potential sacrifice of processing speed.

Fusion at the model-level

Mai et al. [115] pioneered various textual LSTM forms hierarchically, accounting for interdependencies between video texts [118]. Results showcased 80.3%, 68.1%, and 76.1% accuracy in MOSI, MOUD, and IEMOCAP datasets. Moreover, Mehrabian [23] proposed a DL-based multitasking model for sentiment analysis and recognition, factoring in the varying significance of each method through a context-level intermodal attention module (CIM).

Hybrid-level fusion

Zhang et al. [80] combined feature and decision-based strategies to counterbalance their limitations. Utilizing SVM for text-based sentiment classification, they integrated audio-visual indicators using Bi-LSTM for sentiment detection. The culmination of both results through a weighted strategy determined the final sentiment. The fusion of various modalities offers an enriched and holistic approach to emotion and sentiment detection, as the aforementioned studies corroborate.

The architecture of multimedia information entry, including audio, image, and text information, using the most important audio and text feature extraction methods and feature fusion types, is shown in *Fig. 8*.

6|Emotional Computing Databases

In academic research on emotional computing, DBs can be delineated according to the nature of the data they encompass: textual, speech or audio, visual, physiological, and multimodal datasets. The inherent properties of these DBs substantially influence the formulation of models and the architectural design of networks for emotional computation.

6.1|Databases of Multimodal Sentiment

In daily interactions, individuals utilize multimodal signals to convey and comprehend emotions. Multimodal DBs can be broadly categorized into two primary types: non-physical DBs and physical-physiological DBs. An overview of these multimodal DBs is provided in *Table 6*. In emotion analysis, several prominent DBs have been developed to aid research. The Speech Analysis and Interpretation Laboratory introduced the Interactive Emotional Dyadic Motion Capture (IEMOCAP) [122]. This dataset was collected with ten actors who performed scripts designed to elicit various emotional responses, both scripted and improvised. Comprehensive data on facial expressions and hand gestures, encompassing the actors' faces, heads, and

hands, have been documented. The emotions displayed are categorized using two renowned sentiment analysis methods: discrete classifier-based annotation and continuous feature-based annotation.

Furthermore, the CreativeIT DB [122], [123] encompasses multifaceted data sources, including visual, auditory, textual, and full-body motion data. This information was gleaned from sixteen actors engaged in dyadic emotional interactions, each lasting from two to ten minutes. The actors engaged in two spontaneous interactions: binary exercises and recitations. The emotional states of each participant are annotated in a three-dimensional framework following the video's frame rate. Moreover, the HOW DB [123] incorporates a collection of

Table 5. An overview visual-audio-text ER.

Publication	Feature Representation	Classifier	Fusion Strategy	Database	Performance
[80]	Facial movement, MFCC	SVM, BiLSMT	Hybrid-level	SEMAINE	22-class: 80.50, 665.20
[121]	CNN, handcrafted, CFS, PCA	MKL	Feature-level Decision-level	HOW	3 classes: 88.60 3 classes: 86.27
[23]	Three Bi-GRU	CIM-attention	Model-based	CMU-MOSEI	2 Multi-label: 62.80
[85]	Proxy and Attention Multiplicative fusion	FC	Feature-level	IEMOCAP CMU-MOSEI	4 classes1:82.70 6 classes1:89.00
[124]	ResNet-v2	LSTM, fusion, Ensemble model	Hybrid-level	IEMOCAP	4 classes 82.90
[125]	Audio: LLDs Visual: 3D-CNN Text: Word2vec	FC + Softmax	Model-level	CMU-MOSI	Acc(2-class):81.3%
[84]	Audio: LLDs Visual: 3D-CNN Text: Word2vec	FC + Softmax	Hybrid-level	IEMOCAP	Acc(4-class):73.94%
[126]	Audio: LLDs Visual: Facial Landmarks Text: Glove	FC + Softmax	Model-level	1)IEMOCAP 2)CMU-MOSEI	1) Acc(4-class):82.7 F1-score:82.40% 2) Acc(6- class):89.0% F1-score:90.2%
[127]	Audio: LLDs Visual:3D-CNN Text: CNN	FC + Softmax	Model-level	1)IEMOCAP 2)CMU-MOSI 3)MELD	1)Acc(6-class):60.81% 2)Acc(2-class):82.71% 3)Acc(7-class):67.04%
[128]	Audio: CNN Visual: CNN Text: Transformer	FC + Softmax	Decision-level	1)IEMOCAP 2)CMU-MOSEI	1) Acc(6- class):84.4% F1-score:57.4% 2) Acc(6- class):66.8% F1-score:46.8%
[115]	Audio: LLDs Visual: 3D-CNN Text: Glove	FC + Softmax	Model-level	IEMOCAP	Acc(6-class):65.0% F1-score: 64.5%
[129]	Audio:Bi-GRU Visual: Bi-GRU Text: Bi-GRU	FC + Softmax	Model-level	1)IEMOCAP 2)CMU-MOSEI 3)CMU-MOSI 4)MELD	1) Acc(4-class):83.45% F1-score:82.63% 2) Acc(2- class):82.4% Acc(7-class):50.9% , F1-score:82.6% 3): Acc(2-class):82.3% Acc(7-class):39.4% , F1-score:82.5% MAE:0.896 Corr:0.697 4) 0.598 , Corr:0.69

Table 5. Continued.

Publication	Feature Representation	Classifier	Fusion Strategy	Database	Performance
[130]	Audio: AlexNet Visual: ResNet Text: Word2vec	FC + Softmax	Model-level	1)IEMOCAP 2)MSP-IMPROV 3)CMU-MOSI 4)CMU-MOSEI	1) Acc(4-class):86.3 F1-score:86.5% 2) Acc(4- class):71.8%, F1-score:71.8% 3) Acc(2- class):85.2% Acc(7-class):46.6% F1-score:85.1% MAE:0.713 , Corr:0.790 4):Acc(2- class):85.4% Acc(7-class):52.8% , F1-score:85.6% MAE:0.601 , Corr:0.776
[129]	Audio: LLDs Visual: Facial Landmarks Text: Glove	Contrastive Learning	Model-level	1)CMU-MOSI 2)CMU-MOSEI	1)Acc(7-class):48.3% 2)Acc(7-class):53.4%
[131]	Audio: Wav2Vec2.0 Visual: DenseNet Text: BERT	FC + Softmax	Model-level	1)IEMOCAP 2)MSP-IMPROV	1)WA(4- class):80.01% UAR(4-class):81.09% , Acc(4-class):80.01% 2)WA(4- class):72.36% UAR(4-class):72.22% Acc(4-class):72.36%

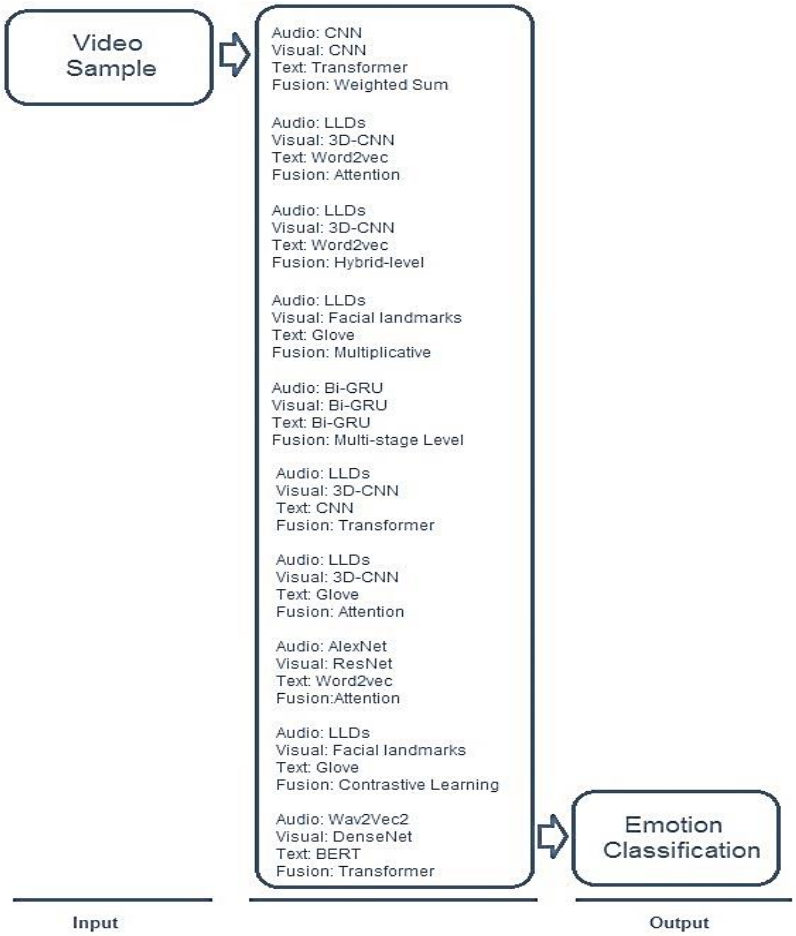


Fig. 8. Here is an overview of the framework for diverse trimodal ER, incorporating trimodal feature extraction and fusion methods.

13 positive, 12 negative, and 22 neutral videos, all sourced from YouTube. Additionally, relevant comments from these videos have been extracted and documented. The ICT-MMMO DB [132] is an expansive collection of 308 YouTube videos complemented by 78 film review clips from ExpoTV. Sentiment in this DB is segmented into five distinct categories: strongly positive, weakly positive, neutral, strongly negative, and weakly negative. Notably, the CMU-MOSEI [133] stands as a preeminent resource for sentiment analysis and recognition, boasting 23,453 sentences and 3,228 video segments acquired from over 1,000 distinct YouTube contributors. Each video in this dataset is paired with a meticulously crafted transcription that synchronizes the audio with corresponding phoneme scores.

Additionally, the MAHNOB-HCI [81] DB is a video-physiological DB. The emotions of 27 participants were monitored and recorded while watching 20 films using 6 video cameras, a head-worn microphone, an eye gaze tracker, and physiological sensors. The Remote Collaborative and Affective Interactions (RECOLA) [134] is a multimodal corpus of spontaneous interactions from 46 participants (in French). The participants worked in pairs to discuss a disaster scenario escape plan and reached an agreement via remote video conferencing. The participants' activities are recorded and annotated by six annotators with two continuous emotional dimensions, arousal and valence, as well as social behavior labels on five dimensions. DECAF [135] is a DB based on Magnetoencephalography for decoding the affective responses of 30 subjects while watching 36 movie clips and 40 one-minute music video clips. DECAF provides a detailed analysis of the correlations between the recorded data.

Table 6. An overview of these multimodal DBs.

Name	Components				Subjects	Type	Emotion Categories
	Text	Speech	Visual	Psych			
IEMOCAP [122]	✓	✓	✓	×	10	Acted	Happiness, Anger, Sad, Frustration and Neutral Activation-Valence-Dominance
CreativeIT [136], [137]	✓	✓	✓	×	16	Induced	Activation-Valence-Dominance
HOW [138]	✓	✓	✓	×	×	Natural	Positive, Negative and Neutral
ICT-MMMO [132]	✓	✓	✓	×	×	Natural	Five basic sentiments
CMU-MOSEI [113]	✓	✓	✓	×	×	Natural	Six basic emotions, Five basic sentiments
MAHNOB-HC [112]	×	✓	✓	✓	27	Induced	Arousal-Valence-Dominance-Predictability Disgust, Amusement, Joy, Fear, Sadness, Neutral
RECOLA [134]	×	✓	✓	✓	46	Natural	Arousal-Valence, Agreement, Dominance, Engagement, Performance and Rapport
DECAF [135]	×	✓	✓	✓	30	Induced	Arousal-Valence-Dominance, Six basic expressions Amusing, Funny and Exciting

7 | Evaluation Metrics

In any growing field of research, it is necessary to establish a commonly accepted evaluation methodology widely used within the field. This holds for sentiment classification as well. At present, the majority of the studies surveyed adopt the following standard measures:

I. Accuracy: Accuracy measures the number of correct predictions divided by the total number of predictions. It is calculated by taking the true positives and negatives ratio to the total number of samples. Mathematically, the accuracy, ACC, is given by

$$\mathrm{ACC} = \frac{\mathrm{TN} + \mathrm{TP}}{\mathrm{TN} + \mathrm{TP} + \mathrm{FP} + \mathrm{FN}}, \quad (1)$$

where TN is the tally of true negatives, TP is the tally of true positives, FP is the tally of false positives, and FN is the tally of false negatives.

II. Precision: Precision is a measure of accuracy, indicating the extent to which the model's predictions are correct. It is the proportion of true positives to the total number of positives. Mathematically, precision, PRE, is given by

$$\mathrm{PRE} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}. \quad (2)$$

A lower precision value indicates a higher number of false positives, and a higher one indicates a lower number. Therefore, a higher Precision score indicates better performance.

III. Recall: Recall measures the totality of the model's predictions and indicates the extent to which all relevant samples have been correctly classified. It is the ratio of true positive predictions to the sum of true positive and false negative predictions. Mathematically, the recall, REC, is given by

$$\mathrm{REC} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}. \quad (3)$$

A higher recall indicates that more relevant samples have been correctly classified, while a lower recall indicates many false negatives in the model's predictions.

IV. F-Measure: The F-measure is a metric that balances recall and precision in sentiment categorization. It is calculated as the harmonic mean of precision and recall, given by

$$\mathrm{F1} - \mathrm{score} = \frac{\mathrm{PRE} \times \mathrm{REC}}{\mathrm{PRE} + \mathrm{REC}}. \quad (4)$$

A higher score indicates a better model performance in capturing the true positive cases and avoiding the false positive cases. In other words, a high F1 score means the model is well-balanced between correctly identifying the relevant results and having fewer false positives.

V. K-Fold Cross-Validation: ML models can be evaluated using cross-validation resampling techniques. In k-fold cross-validation, a limited sample of data is resampled and divided into k groups. The model is then trained on a single group and tested on k -1 groups. K-fold cross-validation is used to avoid underfitting and overfitting.

VI. Confusion Matrix: The performance of a model is visualized by a confusion matrix (*Table 7*). The different metrics used to evaluate the model performance can be calculated by analyzing the confusion matrix.

Table 7. Confusion matrix.

	Actual Values		
Predicted Values		Negative	Positive
	Negative	TN	FN
	Positive	FP	TP

8 | Ethical Considerations in Emotional Computing

Emotional computing raises ethical considerations regarding privacy due to collecting and analyzing personal and sensitive data. Concerns exist about the potential manipulation of emotions or behaviors through targeted content and advertising. The potential consequences of emotional computing include the erosion of human autonomy and emotional privacy and the risk of perpetuating biased or discriminatory outcomes. It is crucial to consider the impact on mental health and well-being when relying on emotional computing, as it may lead to emotional dependency or detachment. Additionally, ethical concerns arise regarding transparency and consent in using emotional data and the potential for unethical or malicious applications, such as emotional manipulation or surveillance. Balancing technological advancement with ethical responsibility is crucial in

developing and deploying emotional computing technologies. Emotional computing also raises questions about the ownership and control of emotional data, particularly in cases where individuals may not fully comprehend the implications of sharing their emotional responses with technology. The importance of robust data protection and user consent mechanisms to safeguard against exploitation and misuse of emotional data is highlighted. Additionally, emotional computing has the potential to perpetuate social inequalities through biased algorithms or discriminatory practices, disproportionately affecting certain groups or individuals. Addressing ethical considerations is crucial for building trust, promoting responsible innovation, and ensuring that emotional computing technologies serve the greater good while respecting individual autonomy and emotional well-being.

9 | Discussion

This scholarly examination delineates prevalent models and DBs instrumental for emotion computation, focusing on MER and analysis. Specifically, we will explore:

I. The influence of varied signals such as text, audio, and visual cues and the consequent ramifications of their combinatory and fusion methodologies on multimodal sentiment analysis, as evidenced in references [37], [110], [113], [115].

II. The repercussions of certain determinants, for instance, established DBs and performance metrics, upon emotion computational processes.

III. Practical implementations of emotional computing in real-world contexts.

9.1 | Model Combination and Fusion Strategies in Multimodal Sentiment Analysis

Integrating diverse techniques and employing fusion strategies are pivotal elements in multimodal sentiment analysis. Multimodal combinations can be categorized into non-physical connection methods [139], [140], physiological methods [141], and hybrid approaches [142]. Fusion strategies encompass feature-level fusion [91], decision-level fusion [36], combination-level fusion [80], and model-level fusion. Distinct method combinations that do not necessitate physical connections include video-audio, text-audio, and tri-modal video-audio-text configurations [143]. Visual and auditory information integration often results in augmented performance relative to single-mode analyses [144]. Text-based methodologies are central in multimodal sentiment analysis [85], [121]. In contrast, multimodal physiological signals can be integrated with visual cues such as facial expressions, voice, and gestures for comprehensive emotion assessment [145]. At the heart of multimodal sentiment analysis lie two cardinal fusion strategies: 1) feature-level, and 2) decision-level fusion [36], [91]. Feature-level fusion typically involves vector concatenation [85] or fusion [93].

Meanwhile, decision-level fusion often employs majority or average voting mechanisms. Linear-weighted computations [109] serve both feature-level combinations and decision-making, leveraging diverse methodologies. Evidently, feature-level fusion dominates over decision-level fusion in prevalence. However, individual models are processed independently in decision-level fusion, subsequently amalgamating results. Conversely, combination-level fusion [80] seeks to harness advantages from both the previous strategies.

Model-level fusion [146], distinct from the other strategies, utilizes HMM [33] or Bayesian networks [147] to decipher and synergize characteristics from diverse models, the selection of which crucially influences the fusion outcomes.

9.2 | Influence of ML and DL on Emotional Computation

Historically, sentiment analysis predominantly leveraged ML techniques [10], [18], [19]. This process involved signal preprocessing, curated feature extraction, and classification (subject to feature selection) [139], [148]–[151]. Despite their versatility, the specificity of ML features can impede their repurposing. Common classifiers in this domain include SVM [152], HMM [135], GMM [138], RF [133], KNN [134], and ANN

[111], with SVM emerging as particularly efficacious in emotional computations [83], [153]. DL offers distinct advantages, particularly in automated feature acquisition. CNNs and their derivatives excel in distilling salient features from static data, while RNNs capture temporal dynamics in sequences [154]–[159]. CNN-LSTM models undertake deep spatiotemporal feature extraction, with adversarial learning enhancing model robustness through data augmentation [160]–[163]. Additionally, autoencoders, constructed through DL, bolster overall performance [156], [164], [165].

9.3 | Factors Influencing Sentiment Computation

This review underscores the influence of DB metrics on emotion computation progress. Notably, the scarcity of non-contact physiological emotion DBs has stymied video-PER [166]. In contrast, the ascendance of Facial Emotion Recognition (FER) is noteworthy, spurred by publicly accessible DBs. However, DB disparities in size, quality, and collection environments pose challenges, especially affecting DL-based ER more than its ML counterpart. Many findings underline the pivotal role of DB size in sentiment analysis performance [167], [168], with pre-trained DL models [169], [170] proffering solutions for personalized sentiment analyses. Emotion DBs are often tagged with discrete or dimensional labels corresponding to the raw data. Emotional categorization spans discrete emotional states, dimensional classification, and continuous dimensional regression [171], [172]. Notably, F-Measure is the favored evaluation metric, especially for skewed DBs. Concurrently, MSE and RMSE serve as benchmarks for sentiment prediction [161], [173], while the congruence coefficient, blending PCC and MSR, is recommended for baseline evaluations [174], [175].

9.4 | Practical Implications of Emotional Computing

Recent scholarly endeavors have spotlighted the pragmatic applications of emotional computing [176], [177]. SenticNet, under the aegis of Eric Cambria [81], [178], [179], has harnessed sentiment computing [22], [180], [181], insights for diverse sectors from Human-Computer Interaction (HCI) [182] to finance [152] and social media analytics [183], [184]. Sentiment analysis tools are also integrated into recommender systems [185], [186]. VER finds utility in arenas like distance learning [187], [188], player performance optimization [121], [189]–[191], and healthcare [192], including autism interventions [193]. Audio and physiological cues, consistent in emotional representation, are pivotal in diagnosing mood disorders, augmented by wearable technology [194]–[197]. The potential of multimodal sentiment analysis in bolstering the efficacy of unimodal sentiment recognition has spurred extensive research into practical applications, rendering it an intriguing domain of investigation [198].

10 | Conclusion and Recent Developments

This article offers a thorough examination of the existing literature on emotional computing. In Section 1, the text classifies emotional computing and provides illustrative examples. Section 2 presents a synopsis of contemporary reviews in emotional computing. Section 3 delineates emotional models based on prevailing psychological theories, segmenting them into discrete and subsequent models; the latter's output aids in categorizing emotions. These identifications are subsequently assessed through classification or regression against pertinent metrics. Crucially, the evolution of emotional computing necessitates the use of foundational DBs for training and computational frameworks for DL and ML-grounded emotion comprehension. Section 4 examines widely used multivariate DBs in sentiment analysis, given that a majority of sentiment analysis techniques employ these repositories. Section 5 delves into recent innovations in sentiment computing, primarily segmenting them into multimodal sentiment recognition and analysis, which are further partitioned into ML-centric techniques and DL-centric models. Broadly, multimodal sentiment analysis is bifurcated into non-physical multimodal techniques, such as visual-audio, text-audio, and video-audio, in conjunction with varied strategies like feature-level, decision-level, model-level, or hybrid approaches. Section 6 broaches critical considerations in sentiment computing, emphasizing the ramifications of model integrations and amalgamation strategies on multimodal sentiment analysis and the implications of ML and DL methodologies. While significant strides have been made in crafting computational systems utilizing unimodal or multimodal

data, resilient and productive algorithms for emotion prediction and discernment in multifaceted environments remain scarce. Consequently, this article proposes several imperative directives for prospective investigations in emotional computing:

I. The creation of innovative and expansive foundational DBs, particularly multimodal repositories encompassing varied modalities such as text, audio, imagery, and physiological metrics. These should encapsulate spontaneous and staged scenarios, ensuring data annotation for discrete and dimensional emotional models.

II. Numerous impediments in emotion detection and analysis await resolution, encompassing facial ER in the face of partial occlusions or insincere expressions, physiological emotion discernment from intricate signals, and a foundational model tailor-made for identifying discrete and subsequent emotional states.

III. Augmenting fusion strategies remains a vast domain for exploration. Incorporating rule-based or statistical expertise can refine the fusion of diverse methodologies, taking into consideration the significance and role of each in ensuring precision in a prospective area for future inquiry.

IV. The employment of unsupervised learning techniques, such as cluster-based learning, warrants further investigation due to their potential to enhance the resilience and consistency of sentiment analysis.

V. Robotics stands out as a salient application of sentiment analysis. With the developments highlighted in this review, one can envision robots endowed with emotional acuity, adept at emulating human interactions and surroundings, and swiftly delivering suitable reactions.

An area with promising prospects for future research is the comprehension of sentiment in conversations. Expressions of sentiment by one individual in a conversation can influence others. Previous studies have shown the significance of discourse context in understanding human language. If multimodal systems can replicate human emotional interdependencies, it could substantially advance multimodal sentiment analysis. Moreover, there is a need for further research to develop models that are language-independent, enabling generalization to prediction tasks across different languages.

Conflict of Interest

The authors declare no conflict of interest.

Data Availability

All data are included in the text.

Funding

This research received no specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

[1] Fleckenstein, K. S. (1991). Defining affect in relation to cognition: A response to Susan McLeod. *Journal of advanced composition*, *11*(2), 447–453.

[2] Picard, R. W. (1997). *Affective computing*. MIT Press.

[3] Picard, R. W., Vyzas, E., & Healey, J. (2001). Toward machine emotional intelligence: Analysis of affective physiological state. *IEEE transactions on pattern analysis and machine intelligence*, *23*(10), 1175–1191. DOI:10.1109/34.954607

[4] Park, J. S., Kim, J. H., & Oh, Y. H. (2009). Feature vector classification based speech emotion recognition for service robots. *IEEE transactions on consumer electronics*, *55*(3), 1590–1596. DOI:10.1109/TCE.2009.5278031

[5] Scheutz, M. (2012). The affect dilemma for artificial agents: Should we develop affective artificial agents? *IEEE transactions on affective computing*, *3*(4), 424–433. DOI:10.1109/T-AFFC.2012.29

[6] McColl, D., Hong, A., Hatakeyama, N., Nejat, G., & Benhabib, B. (2016). A survey of autonomous human affect detection methods for social robots engaged in natural HRI. *Journal of intelligent and robotic systems: theory and applications*, *82*(1), 101–133. DOI:10.1007/s10846-015-0259-2

[7] Healey, J. A., & Picard, R. W. (2005). Detecting stress during real-world driving tasks using physiological sensors. *IEEE transactions on intelligent transportation systems*, *6*(2), 156–166. DOI:10.1109/TITS.2005.848368

[8] Langeroudi, M. K., Yamaghani, M. R., & Khodaparast, S. (2022). FD-LSTM: a fuzzy LSTM model for chaotic time-series prediction. *IEEE intelligent systems*, *37*(4), 70–78. DOI:10.1109/MIS.2022.3179843

[9] Cambria, E. (2016). Affective Computing and Sentiment Analysis. *IEEE intelligent systems*, *31*(2), 102–107. DOI:10.1109/MIS.2016.31

[10] Munezero, M., Montero, C. S., Sutinen, E., & Pajunen, J. (2014). Are they different? affect, feeling, emotion, sentiment, and opinion detection in text. *IEEE transactions on affective computing*, *5*(2), 101–111. DOI:10.1109/TAFFC.2014.2317187

[11] Poria, S., Cambria, E., Bajpai, R., & Hussain, A. (2017). A review of affective computing: From unimodal analysis to multimodal fusion. *Information fusion*, *37*, 98–125. DOI:10.1016/j.inffus.2017.02.003

[12] Rouast, P. V., Adam, M. T. P., & Chiong, R. (2021). Deep Learning for Human Affect Recognition: Insights and New Developments. *IEEE transactions on affective computing*, *12*(2), 524–543. DOI:10.1109/TAFFC.2018.2890471

[13] Shoumy, N. J., Ang, L. M., Seng, K. P., Rahaman, D. M. M., & Zia, T. (2020). Multimodal big data affective analytics: A comprehensive survey using text, audio, visual and physiological signals. *Journal of network and computer applications*, *149*, 102447. DOI:10.1016/j.jnca.2019.102447

[14] Ekman, P. (1999). Basic emotions. *Handbook of cognition and emotion*, *98*, 45–60. DOI:10.1002/0470013494.ch3

[15] Mehrabian, A. (1980). *Basic dimensions for a general psychological theory: Implications for personality, social, environmental, and developmental studies*. Gunn & Hain Cambridge.

[16] Jiang, Y., Li, W., Hossain, M. S., Chen, M., Alelaiwi, A., & Al-Hammadi, M. (2020). A snapshot research and implementation of multimodal information fusion for data-driven emotion recognition. *Information fusion*, *53*, 209–221. DOI:10.1016/j.inffus.2019.06.019

[17] Corneanu, C. A., Simón, M. O., Cohn, J. F., & Guerrero, S. E. (2016). Survey on RGB, 3D, thermal, and multimodal approaches for facial expression recognition: history, trends, and affect-related applications. *IEEE transactions on pattern analysis and machine intelligence*, *38*(8), 1548–1568. DOI:10.1109/TPAMI.2016.2515606

[18] El Ayadi, M., Kamel, M. S., & Karray, F. (2011). Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern recognition*, *44*(3), 572–587. DOI:10.1016/j.patcog.2010.09.020

[19] Zhang, J., Yin, Z., Chen, P., & Nichele, S. (2020). Emotion recognition using multi-modal data and machine learning techniques: A tutorial and review. *Information fusion*, *59*, 103–126. DOI:10.1016/j.inffus.2020.01.011

[20] Poria, S., Cambria, E., & Gelbukh, A. (2016). Aspect extraction for opinion mining with a deep convolutional neural network. *Knowledge-based systems*, *108*, 42–49. DOI:10.1016/j.knosys.2016.06.009

[21] Cambria, E., Speer, R., Havasi, C., & Hussain, A. (2010). *SenticNet: a publicly available semantic resource for opinion mining* [presentation]. AAAI fall symposium - technical report (Vol. FS-10-02, pp. 14–18). https://cdn.aaai.org/ocs/2216/2216-9491-2-PB.pdf

[22] Akhtar, M. S., Chauhan, D. S., Ghosal, D., Poria, S., Ekbal, A., & Bhattacharyya, P. (2019). *Multi-task learning for multi-modal emotion recognition and sentiment analysis*. https://doi.org/10.48550/arXiv.1905.05812

[23] Mehrabian, A. (2008). *Communication Theory*. Routledge.

[24] Alm, C. O., Roth, D., & Sproat, R. (2005). *Emotions from text: machine learning for text-based emotion prediction* [presentation]. HLT/emnlp 2005 - human language technology conference and conference on empirical methods in natural language processing, proceedings of the conference (pp. 579–586). https://aclanthology.org/H05-1073.pdf

[25] Liu, Z. T., Xie, Q., Wu, M., Cao, W. H., Mei, Y., & Mao, J. W. (2018). Speech emotion recognition based on an improved brain emotion learning model. *Neurocomputing*, *309*, 145–156.

[26] Sajjad, M., Nasir, M., Ullah, F. U. M., Muhammad, K., Sangaiah, A. K., & Baik, S. W. (2019). Raspberry Pi assisted facial expression recognition framework for smart security in law-enforcement services. *Information sciences*, *479*, 416–431. DOI:10.1016/j.ins.2018.07.027

[27] Sarkar, P., & Etemad, A. (2022). Self-supervised ECG representation learning for emotion recognition. *IEEE transactions on affective computing*, *13*(3), 1541–1554. DOI:10.1109/TAFFC.2020.3014842

[28] Alarcão, S. M., & Fonseca, M. J. (2019). Emotions recognition using EEG signals: A survey. *IEEE transactions on affective computing*, *10*(3), 374–393. DOI:10.1109/TAFFC.2017.2714671

[29] Kim, J., & André, E. (2008). Emotion recognition based on physiological changes in music listening. *IEEE transactions on pattern analysis and machine intelligence*, *30*(12), 2067–2083. DOI:10.1109/TPAMI.2008.26

[30] Tzirakis, P., Trigeorgis, G., Nicolaou, M. A., Schuller, B. W., & Zafeiriou, S. (2017). End-to-end multimodal emotion recognition using deep neural networks. *IEEE journal of selected topics in signal processing*, *11*(8), 1301–1309.

[31] Baltrusaitis, T., Robinson, P., & Morency, L. P. (2012). 3D constrained local model for rigid and non-rigid facial tracking. *Proceedings of the ieee computer society conference on computer vision and pattern recognition* (pp. 2610–2617). IEEE. DOI: 10.1109/CVPR.2012.6247980

[32] Lin, J. C., Wu, C. H., & Wei, W. L. (2012). Error weighted semi-coupled hidden markov model for audio-visual emotion recognition. *IEEE transactions on multimedia*, *14*(1), 142–156. DOI:10.1109/TMM.2011.2171334

[33] Huang, X., Kortelainen, J., Zhao, G., Li, X., Moilanen, A., Seppänen, T., & Pietikäinen, M. (2016). Multi-modal emotion analysis from facial expressions and electroencephalogram. *Computer vision and image understanding*, *147*, 114–124. DOI:10.1016/j.cviu.2015.09.015

[34] Wang, Z., Zhou, X., Wang, W., & Liang, C. (2020). Emotion recognition using multimodal deep learning in multiple psychophysiological signals and video. *International journal of machine learning and cybernetics*, *11*(4), 923–934.

[35] Meng, T., Jing, X., Yan, Z., & Pedrycz, W. (2020). A survey on machine learning for data fusion. *Information fusion*, *57*, 115–129.

[36] Zhang, S., Zhang, S., Huang, T., Gao, W., & Tian, Q. (2017). Learning affective features with a hybrid deep model for audio--visual emotion recognition. *IEEE transactions on circuits and systems for video technology*, *28*(10), 3030–3043.

[37] Ko, B. C. (2018). A brief review of facial emotion recognition based on visual information. *Sensors (Switzerland)*, *18*(2), 401. DOI:10.3390/s18020401

[38] Li, S., & Deng, W. (2020). Deep facial expression recognition: a survey. *Journal of image and graphics*, *25*(11), 2306–2320. DOI:10.11834/jig.200233

[39] Merghani, W., Davison, A. K., & Yap, M. H. (2018). *A review on facial micro-expressions analysis: datasets, features and metrics*. http://arxiv.org/abs/1805.02397

[40] Alexandre, G. R., Soares, J. M., & Pereira Thé, G. A. (2020). Systematic review of 3D facial expression recognition methods. *Pattern recognition*, *100*, 107108. DOI:10.1016/j.patcog.2019.107108

[41] Liu, R., Shi, Y., Ji, C., & Jia, M. (2019). A survey of sentiment analysis based on transfer learning. *IEEE access*, *7*, 85401–85412.

[42] Khalil, R. A., Jones, E., Babar, M. I., Jan, T., Zafar, M. H., & Alhussain, T. (2019). Speech emotion recognition using deep learning techniques: A review. *IEEE access*, *7*, 117327–117345.

[43] Patel, K., Mehta, D., Mistry, C., Gupta, R., Tanwar, S., Kumar, N., & Alazab, M. (2020). Facial sentiment analysis using AI techniques: state-of-the-art, taxonomies, and challenges. *IEEE access*, *8*, 90495–90519. DOI:10.1109/ACCESS.2020.2993803

[44] Noroozi, F., Corneanu, C. A., Kaminska, D., Sapinski, T., Escalera, S., & Anbarjafari, G. (2021). Survey on emotional body gesture recognition. *IEEE transactions on affective computing*, *12*(2), 505–523. DOI:10.1109/TAFFC.2018.2874986

[45] Poria, S., Majumder, N., Mihalcea, R., & Hovy, E. (2019). Emotion recognition in conversation: Research challenges, datasets, and recent advances. *IEEE access*, *7*, 100943–100953. DOI:10.1109/ACCESS.2019.2929050

[46] Yue, L., Chen, W., Li, X., Zuo, W., & Yin, M. (2019). A survey of sentiment analysis in social media. *Knowledge and information systems*, *60*, 617–663.

[47] Wang, Z., Ho, S. B., & Cambria, E. (2020). A review of emotion sensing: categorization models and algorithms. *Multimedia tools and applications*, *79*(47/48), 35553–35582. DOI:10.1007/s11042-019-08328-z

[48] Han, J., Zhang, Z., Cummins, N., & Schuller, B. (2019). Adversarial training in affective computing and sentiment analysis: recent advances and perspectives. *IEEE computational intelligence magazine*, *14*(2), 68–81. DOI:10.1109/MCI.2019.2901088

[49] Bota, P. J., Wang, C., Fred, A. L. N., & Placido Da Silva, H. (2019). A review, current challenges, and future possibilities on emotion recognition using machine learning and physiological signals. *IEEE access*, *7*, 140990–141020. DOI:10.1109/ACCESS.2019.2944001

[50] García-Martínez, B., Martínez-Rodrigo, A., Alcaraz, R., & Fernández-Caballero, A. (2021). A review on nonlinear methods using electroencephalographic recordings for emotion recognition. *IEEE transactions on affective computing*, *12*(3), 801–820. DOI:10.1109/TAFFC.2018.2890636

[51] Ekman, P. (1971). *Universals and cultural differences in facial expressions of emotion.* [presentation]. Nebraska symposium on motivation (pp. 207–283). https://psycnet.apa.org/record/1973-11154-001

[52] Dang, F., Chen, H., & Zhang, P. (2022). Dpt-fsnet: dual-path transformer based full-band and sub-band fusion network for speech enhancement. *ICASSP, IEEE international conference on acoustics, speech and signal processing - proceedings* (Vol. 2022-May, pp. 6857–6861). IEEE. DOI: 10.1109/ICASSP43922.2022.9746171

[53] Tracy, J. L., & Randles, D. (2011). Four models of basic emotions: A review of Ekman and Cordaro, Izard, Levenson, and Panksepp and Watt. *Emotion review*, *3*(4), 397–405. DOI:10.1177/1754073911410747

[54] Russell, J. A. (1980). A circumplex model of affect. *Journal of personality and social psychology*, *39*(6), 1161–1178. DOI:10.1037/h0077714

[55] Plutchik, R. (2003). *Emotions and life: Perspectives from psychology, biology, and evolution.* American Psychological Association.

[56] Cambria, E., Livingstone, A., & Hussain, A. (2012). The hourglass of emotions. *Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics)* (Vol. 7403 LNCS, pp. 144–157). Springer. DOI: 10.1007/978-3-642-34584-5_11

[57] Susanto, Y., Livingstone, A. G., Ng, B. C., & Cambria, E. (2020). The hourglass model revisited. *IEEE intelligent systems*, *35*(5), 96–102. DOI:10.1109/MIS.2020.2992799

[58] Lopes, A. T., de Aguiar, E., De Souza, A. F., & Oliveira-Santos, T. (2017). Facial expression recognition with Convolutional neural networks: coping with few data and the training sample order. *Pattern recognition*, *61*, 610–628. DOI:10.1016/j.patcog.2016.07.026

[59] Ren, Z., Baird, A., Han, J., Zhang, Z., & Schuller, B. (2020). Generating and protecting against adversarial attacks for deep speech-based emotion recognition models. *ICASSP, IEEE international conference on acoustics, speech and signal processing - proceedings* (Vol. 2020, pp. 7184–7188). IEEE. DOI: 10.1109/ICASSP40776.2020.9054087

[60] Wang, Z., Ho, S. B., & Cambria, E. (2020). Multi-level fine-scaled sentiment sensing with ambivalence handling. *International journal of uncertainty, fuzziness and knowldege-based systems*, *28*(4), 683–697. DOI:10.1142/S0218488520500294

[61] Bakker, I., van der Voordt, T., Vink, P., & de Boon, J. (2014). Pleasure, arousal, dominance: Mehrabian and Russell revisited. *Current psychology*, *33*(3), 405–421. DOI:10.1007/s12144-014-9219-4

[62] Russell, J. A., & Mehrabian, A. (1977). Evidence for a three-factor theory of emotions. *Journal of research in personality*, *11*(3), 273–294. DOI:10.1016/0092-6566(77)90037-X

[63] Dabas, H., Sethi, C., Dua, C., Dalawat, M., & Sethia, D. (2018). *Emotion classification using eeg signals* [presentation]. Proceedings of the 2018 2nd international conference on computer science and artificial intelligence (pp. 380–384). https://doi.org/10.1145/3297156.3297177

[64] Liang, J., Chen, S., Zhao, J., Jin, Q., Liu, H., & Lu, L. (2019). Cross-culture multimodal emotion recognition with adversarial learning. *ICASSP, IEEE international conference on acoustics, speech and signal processing - proceedings* (Vol. 2019-May, pp. 4000–4004). IEEE. DOI: 10.1109/ICASSP.2019.8683725

[65] Zhang, L., Walter, S., Ma, X., Werner, P., Al-Hamadi, A., Traue, H. C., & Gruss, S. (2017). "BioVid emo db": a multimodal database for emotion analyses validated by subjective ratings. *2016 IEEE symposium series on computational intelligence, ssci 2016* (pp. 1–6). IEEE. DOI: 10.1109/SSCI.2016.7849931

[66] Gu, X., Shen, Y., & Xu, J. (2021). Multimodal emotion recognition in deep learning:a survey. *Proceedings - 2021 international conference on culture-oriented science and technology, ICCST 2021* (pp. 77–82). IEEE. DOI: 10.1109/ICCST53801.2021.00027

[67] Caridakis, G., Castellano, G., Kessous, L., Raouzaiou, A., Malatesta, L., Asteriadis, S., & Karpouzis, K. (2007). Multimodal emotion recognition from expressive faces, body gestures and speech. *IFIP international federation for information processing* (Vol. 247, pp. 375–388). Springer. DOI: 10.1007/978-0-387-74161-1_41

[68] Niu, Z., Zhong, G., & Yu, H. (2021). A review on the attention mechanism of deep learning. *Neurocomputing*, *452*, 48–62. DOI:10.1016/j.neucom.2021.03.091

[69] Chorowski, J., Bahdanau, D., Cho, K., & Bengio, Y. (2014). *End-to-end continuous speech recognition using attention-based recurrent NN: first results*. http://arxiv.org/abs/1412.1602

[70] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., … Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems* (Vol. 2017-Decem, pp. 5999–6009). NIPS.

[71] Khan, S., Naseer, M., Hayat, M., Zamir, S. W., Khan, F. S., & Shah, M. (2022). Transformers in Vision: A Survey. *ACM computing surveys*, *54*(10), 1–41. DOI:10.1145/3505244

[72] Gong, Y., Lai, C. I. J., Chung, Y. A., & Glass, J. (2022). *SSAST: self-supervised audio spectrogram transformer* [presentation]. Proceedings of the 36th AAAI conference on artificial intelligence (Vol. 36, pp. 10699–10709). DOI: 10.1609/aaai.v36i10.21315

[73] Selva, J., Johansen, A. S., Escalera, S., Nasrollahi, K., Moeslund, T. B., & Clapés, A. (2023). Video transformers: A survey. *IEEE transactions on pattern analysis and machine intelligence*. https://ieeexplore.ieee.org/abstract/document/10041724/

[74] Sun, L., Zhao, G., Zheng, Y., & Wu, Z. (2022). Spectral-spatial feature tokenization transformer for hyperspectral image classification. *IEEE transactions on geoscience and remote sensing*, *60*, 1–14. DOI:10.1109/TGRS.2022.3144158

[75] Dai, L., Liu, H., Tang, H., Wu, Z., & Song, P. (2023). AO2-DETR: arbitrary-oriented object detection transformer. *IEEE transactions on circuits and systems for video technology*, *33*(5), 2342–2356. DOI:10.1109/TCSVT.2022.3222906

[76] Zhang, S., Zhang, X., Zhao, X., Fang, J., Niu, M., Zhao, Z., … Tian, Q. (2023). MTDAN: A lightweight multi-scale temporal difference attention networks for automated video depression detection. *IEEE transactions on affective computing*, 1-12. DOI:10.1109/TAFFC.2023.3312263

[77] Zhang, Z., & Zhang, S. (2023). Modeling air quality PM2.5 forecasting using deep sparse attention-based transformer networks. *International journal of environmental science and technology*, *20*(12), 13535–13550. DOI:10.1007/s13762-023-04900-1

[78] Sarkar, C., Bhatia, S., Agarwal, A., & Li, J. (2014). *Feature analysis for computational personality recognition using youtube personality data set* [presentation]. WCPR 2014 - proceedings of the 2014 workshop on computational personality recognition, workshop of mm 2014 (pp. 11–14). DOI: 10.1145/2659522.2659528

[79] Verma, G. K., & Tiwary, U. S. (2014). Multimodal fusion framework: A multiresolution approach for emotion classification and recognition from physiological signals. *NeuroImage*, *102*(P1), 162–172. DOI:10.1016/j.neuroimage.2013.11.007

[80] Zhang, X., Liu, J., Shen, J., Li, S., Hou, K., Hu, B., … Zhang, T. (2021). Emotion recognition from multimodal physiological signals using a regularized deep fusion of kernel machine. *IEEE transactions on cybernetics*, *51*(9), 4386–4399. DOI:10.1109/TCYB.2020.2987575

[81] Chen, J., Chen, Z., Chi, Z., & Fu, H. (2016). Facial expression recognition in video with multiple feature fusion. *IEEE transactions on affective computing*, *9*(1), 38–50.

[82] Poria, S., Chaturvedi, I., Cambria, E., & Hussain, A. (2016). Convolutional mkl based multimodal emotion recognition and sentiment analysis. *Proceedings - IEEE international conference on data mining, Icdm* (Vol. 0, pp. 439–448). IEEE. DOI: 10.1109/ICDM.2016.178

[83] Do, H. H., Prasad, P. W. C., Maag, A., & Alsadoon, A. (2019). Deep learning for aspect-based sentiment analysis: a comparative review. *Expert systems with applications*, *118*, 272–299. DOI:10.1016/j.eswa.2018.10.003

[84] Mittal, T., Bhattacharya, U., Chandra, R., Bera, A., & Manocha, D. (2020). *M3ER: multiplicative multimodal emotion recognition using facial, textual, and speech cues* [presentation]. AAAI 2020 - 34th AAAI conference on artificial intelligence (Vol. 34, pp. 1359–1367). DOI: 10.1609/aaai.v34i02.5492

[85] Lin, Y., Ji, P., Chen, X., & He, Z. (2023). Lifelong Text-audio sentiment analysis learning. *Neural networks*, *162*, 162–174. DOI:10.1016/j.neunet.2023.02.008

[86] Eskimez, S. E., Maddox, R. K., Xu, C., & Duan, Z. (2020). Noise-resilient training method for face landmark generation from speech. *IEEE/ACM transactions on audio speech and language processing*, *28*, 27–38. DOI:10.1109/TASLP.2019.2947741

[87] Song, M., Bu, J., Chen, C., & Li, N. (2004). Audio-visual based emotion recognition-a new approach [presentation]. *Proceedings of the IEEE computer society conference on computer vision and pattern recognition* (Vol. 2, pp. II--II). IEEE. DOI: 10.1109/cvpr.2004.1315276

[88] Nickel, K., Gehrig, T., Stiefelhagen, R., & McDonough, J. (2005). *A joint particle filter for audio-visual speaker tracking* [presentation]. Proceedings of the seventh international conference on multimodal interfaces, ICMI'05 (pp. 61–68). DOI: 10.1145/1088463.1088477

[89] Zeng, Z., Hu, Y., Liu, M., Fu, Y., & Huang, T. S. (2006). *Training combination strategy of multi-stream fused hidden markov model for audio-visual affect recognition* [presentation]. Proceedings of the 14th annual ACM international conference on multimedia, mm 2006 (pp. 65–68). DOI: 10.1145/1180639.1180661

[90] Caridakis, G., Malatesta, L., Kessous, L., Amir, N., Raouzaiou, A., & Karpouzis, K. (2006). Modeling naturalistic affective states via facial and vocal expressions recognition. *Proceedings of the 8th international conference on multimodal interfaces* (pp. 146–154). Association for Computing Machinery.

[91] Zhao, S., Ma, Y., Gu, Y., Yang, J., Xing, T., Xu, P., … Keutzer, K. (2020). *An end-to-end visual-audio attention network for emotion recognition in user-generated videos* [presentation]. AAAI 2020 - 34th AAAI conference on artificial intelligence (Vol. 34, pp. 303–311). DOI: 10.1609/aaai.v34i01.5364

[92] Zhang, Y., Wang, Z. R., & Du, J. (2019, July). Deep fusion: An attention guided factorized bilinear pooling for audio-video emotion recognition. *2019 international joint conference on neural networks (IJCNN)* (pp. 1-8). IEEE.

[93] Hara, K., Kataoka, H., & Satoh, Y. (2018). *Can spatiotemporal 3d CNNS retrace the history of 2d CNNS and imagenet?* [presentation]. Proceedings of the IEEE computer society conference on computer vision and pattern recognition (pp. 6546–6555). DOI: 10.1109/CVPR.2018.00685

[94] Priyasad, D., Fernando, T., Denman, S., Sridharan, S., & Fookes, C. (2020). Attention driven fusion for multi-modal emotion recognition. *ICASSP, IEEE international conference on acoustics, speech and signal processing - proceedings* (Vol. 2020-May, pp. 3227–3231). IEEE. DOI: 10.1109/ICASSP40776.2020.9054441

[95] Hosseini, S. S., Yamaghani, M. R., & Poorzaker Arabani, S. (2023). Multimodal modelling of human emotion using sound, image and text fusion. *Signal, image and video processing*, *18*(1), 71–79. DOI:10.1007/s11760-023-02707-8

[96] Wollmer, M., Weninger, F., Knaup, T., Schuller, B., Sun, C., Sagae, K., & Morency, L. P. (2013). You tube movie reviews: Sentiment analysis in an audio-visual context. *IEEE intelligent systems*, *28*(3), 46–53. DOI:10.1109/MIS.2013.34

[97] Nguyen, D., Nguyen, K., Sridharan, S., Ghasemi, A., Dean, D., & Fookes, C. (2017). Deep spatio-temporal features for multimodal emotion recognition. *Proceedings - 2017 IEEE winter conference on applications of computer vision, wacv 2017* (pp. 1215–1223). IEEE. DOI: 10.1109/WACV.2017.140

[98] Krishna, D. N., & Patil, A. (2020). *Multimodal emotion recognition using cross-modal attention and 1d convolutional neural networks* [presentation]. Proceedings of the annual conference of the international speech communication association, interspeech (Vol. 2020-Octob, pp. 4243–4247). DOI: 10.21437/Interspeech.2020-1190

[99] Shahrestani, M. R., Motamed, S., & Yamaghani, M. (2023). Recognition of facial and vocal emotions expressions by SOAR model. *Journal of information systems and telecommunication*, *11*(3), 209–221. DOI:10.61186/jist.39828.11.43.209

[100] Hazarika, D., Gorantla, S., Poria, S., & Zimmermann, R. (2018). Self-attentive feature-level fusion for multimodal emotion detection. *Proceedings - IEEE 1st conference on multimedia information processing and retrieval, MIPR 2018* (pp. 196–201). IEEE. DOI: 10.1109/MIPR.2018.00043

[101] Huang, J., Tao, J., Liu, B., Lian, Z., & Niu, M. (2020, May). Multimodal transformer fusion for continuous emotion recognition. *ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 3507-3511). IEEE. DOI: 10.1109/ICASSP40776.2020.9053762

[102] Liu, J., Chen, S., Wang, L., Liu, Z., Fu, Y., Guo, L., & Dang, J. (2021). Multimodal emotion recognition with capsule graph convolutional based representation fusion. *ICASSP 2021-2021 IEEE International conference on acoustics, speech and signal processing (ICASSP)* (pp. 6339-6343). IEEE.

[103] Zhou, H., Du, J., Zhang, Y., Wang, Q., Liu, Q. F., & Lee, C. H. (2021). Information fusion in attention networks using adaptive and multi-level factorized bilinear pooling for audio-visual emotion recognition. *IEEE/ACM transactions on audio speech and language processing*, *29*, 2617–2629. DOI:10.1109/TASLP.2021.3096037

[104] Middya, A. I., Nag, B., & Roy, S. (2022). Deep learning based multimodal emotion recognition using model-level fusion of audio–visual modalities. *Knowledge-based systems*, *244*, 108580. DOI:10.1016/j.knosys.2022.108580

[105] Fu, Y., Okada, S., Wang, L., Guo, L., Song, Y., Liu, J., & Dang, J. (2022). Context- and knowledge-aware graph convolutional network for multimodal emotion recognition. *IEEE multimedia*, *29*(3), 91–100. DOI:10.1109/MMUL.2022.3173430

[106] Yoon, S., Byun, S., & Jung, K. (2018). Multimodal speech emotion recognition using audio and text. *2018 IEEE spoken language technology workshop, slt 2018 - proceedings* (pp. 112–118). IEEE. DOI: 10.1109/SLT.2018.8639583

[107] Cai, L., Hu, Y., Dong, J., & Zhou, S. (2019). Audio-textual emotion recognition based on improved neural networks. *Mathematical problems in engineering*, *2019*, 1–9. DOI:10.1155/2019/2593036

[108] Wu, C. H., & Liang, W. B. (2010). Emotion recognition of affective speech based on multiple classifiers using acoustic-prosodic information and semantic labels. *IEEE transactions on affective computing*, *2*(1), 10-21.

[109] Jin, Q., Li, C., Chen, S., & Wu, H. (2015). Speech emotion recognition with acoustic and lexical features. *ICASSP, IEEE international conference on acoustics, speech and signal processing - proceedings* (Vol. 2015, pp. 4749–4753). IEEE. DOI: 10.1109/ICASSP.2015.7178872

[110] Pepino, L., Riera, P., Ferrer, L., & Gravano, A. (2020). Fusion approaches for emotion recognition from speech using acoustic and text-based features. *ICASSP, IEEE international conference on acoustics, speech and signal processing - proceedings* (Vol. 2020-May, pp. 6484–6488). IEEE. DOI: 10.1109/ICASSP40776.2020.9054709

[111] Lian, Z., Liu, B., & Tao, J. (2021). CTNet: conversational transformer network for emotion recognition. *IEEE/acm transactions on audio speech and language processing*, *29*, 985–1000. DOI:10.1109/TASLP.2021.3049898

[112] Zhang, Z., Zhang, S., Zhao, X., Chen, L., & Yao, J. (2022). Temporal difference-based graph transformer networks for air quality PM2.5 prediction: a case study in China. *Frontiers in environmental science*, *10*, 924986. DOI:10.3389/fenvs.2022.924986

[113] Poria, S., Cambria, E., & Gelbukh, A. (2015). *Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis* [presentation]. Conference proceedings - emnlp 2015: conference on empirical methods in natural language processing (pp. 2539–2544). DOI: 10.18653/v1/d15-1303

[114] Pan, Z., Luo, Z., Yang, J., & Li, H. (2020). *Multi-modal attention for speech emotion recognition* [presentation]. Proceedings of the annual conference of the international speech communication association, interspeech (Vol. 2020, pp. 364–368). DOI: 10.21437/Interspeech.2020-1653

[115] Mai, S., Hu, H., Xu, J., & Xing, S. (2022). Multi-Fusion residual memory network for multimodal human sentiment comprehension. *IEEE transactions on affective computing*, *13*(1), 320–334. DOI:10.1109/TAFFC.2020.3000510

[116] Metallinou, A., Wollmer, M., Katsamanis, A., Eyben, F., Schuller, B., & Narayanan, S. (2012). Context-sensitive learning for enhanced audiovisual emotion classification. *IEEE transactions on affective computing*, *3*(2), 184–198.

[117] Cambria, E., Hazarika, D., Poria, S., Hussain, A., & Subramanyam, R. B. V. (2018). Benchmarking multimodal sentiment analysis. *Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics)* (Vol. 10762 LNCS, pp. 166–179). Springer. DOI: 10.1007/978-3-319-77116-8_13

[118] Perez Rosas, V., Mihalcea, R., & Morency, L. P. (2013). Multimodal sentiment analysis of spanish online videos. *IEEE intelligent systems*, *28*(3), 38–45. DOI:10.1109/MIS.2013.9

[119] Arguello, J., & Rosé, C. (2006). *Topic-segmentation of dialogue* [presentation]. Proceedings of the analyzing conversations in text and speech (pp. 42–49). https://www.aclweb.org/anthology/W06-3407

[120] Peng, H., Ma, Y., Poria, S., Li, Y., & Cambria, E. (2021). Phonetic-enriched text representation for Chinese sentiment analysis with reinforcement learning. *Information fusion*, *70*, 88–99. DOI:10.1016/j.inffus.2021.01.005

[121] Savva, N., Scarinzi, A., & Bianchi-Berthouze, N. (2012). Continuous recognition of player's affective body expression as dynamic quality of aesthetic experience. *IEEE transactions on computational intelligence and ai in games*, *4*(3), 199–212. DOI:10.1109/TCIAIG.2012.2202663

[122] Metallinou, A., Lee, C. C., Busso, C., Carnicke, S., & Narayanan, S. (2010). The usc creative it database: a multimodal database of theatrical improvisation. *Proceeding multimodal corpora: advances in capturing, coding and analyzing multimodality* (pp. 1–4). Citeseer.

[123] Morency, L. P., Mihalcea, R., & Doshi, P. (2011). *Towards multimodal sentiment analysis: harvesting opinions from the web* [presentation]. ICMI'11 - proceedings of the 2011 acm international conference on multimodal interaction (pp. 169–176). DOI: 10.1145/2070481.2070509

[124] Poria, S., Cambria, E., Hazarika, D., Majumder, N., Zadeh, A., & Morency, L. P. (2017). *Context-dependent sentiment analysis in user-generated videos* [presentation]. Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: long papers) (pp. 873–883). https://aclanthology.org/P17-1081/

[125] Peng, S., Cao, L., Zhou, Y., Ouyang, Z., Yang, A., Li, X., … Yu, S. (2022). A survey on deep learning for textual emotion analysis in social networks. *Digital communications and networks*, *8*(5), 745–762. DOI:10.1016/j.dcan.2021.10.003

[126] Wang, Z., Wan, Z., & Wan, X. (2020). *TransModality: an end2end fusion method with transformer for multimodal sentiment analysis* [presentation]. The web conference 2020 - proceedings of the world wide web conference, www 2020 (pp. 2514–2520). DOI: 10.1145/3366423.3380000

[127] Dai, W., Cahyawijaya, S., Liu, Z., & Fung, P. (2021). *Multimodal end-to-end sparse model for emotion recognition*. https://doi.org/10.48550/arXiv.2103.09666

[128] Rish I. (2001). *An empirical study of the naive bayes classifier* [presentation]. IJCAI 2001 workshop on empirical methods in artificial intelligence (Vol. 3, pp. 41–46). http://www.cc.gatech.edu/home/isbell/classes/reading/papers/Rish.pdf

[129] Zheng, J., Zhang, S., Wang, Z., Wang, X., & Zeng, Z. (2023). Multi-channel weight-sharing autoencoder based on cascade multi-head attention for multimodal emotion recognition. *IEEE transactions on multimedia*, *25*, 2213–2225. DOI:10.1109/TMM.2022.3144885

[130] Mai, S., Zeng, Y., Zheng, S., & Hu, H. (2023). Hybrid contrastive learning of tri-modal representation for multimodal sentiment analysis. *IEEE transactions on affective computing*, *14*(3), 2276–2289. DOI:10.1109/TAFFC.2022.3172360

[131] Busso, C., Bulut, M., Lee, C. C., Kazemzadeh, A., Mower, E., Kim, S., … Narayanan, S. S. (2008). IEMOCAP: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, *42*(4), 335–359. DOI:10.1007/s10579-008-9076-6

[132] Bagher Zadeh, A., Liang, P. P., Vanbriesen, J., Poria, S., Tong, E., Cambria, E., … Morency, L. P. (2018). *Multimodal language analysis in the wild: cmu-mosei dataset and interpretable dynamic fusion graph* [presentation]. ACL 2018 - 56th annual meeting of the association for computational linguistics, proceedings of the conference (long papers) (Vol. 1, pp. 2236–2246). DOI: 10.18653/v1/p18-1208

[133] Santhoshkumar, R., & Kalaiselvi Geetha, M. (2020). *Human emotion recognition using body expressive feature* [presentation]. Microservices in big data analytics (pp. 141–149). DOI: 10.1007/978-981-15-0128-9_13

[134] Bong, S. Z., Murugappan, M., & Yaacob, S. (2012). Analysis of electrocardiogram (ecg) signals for human emotional stress classification. *Communications in computer and information science* (Vol. 330 CCIS, pp. 198–205). Springer. DOI: 10.1007/978-3-642-35197-6_22

[135] New, T. L., Foo, S. W., & De Silva, L. C. (2003). Detection of stress and emotion in speech using traditional and fft based log energy features. *ICICS-PCM 2003 - proceedings of the 2003 joint conference of the 4th international*

conference on information, communications and signal processing and 4th pacific-rim conference on multimedia (Vol. 3, pp. 1619–1623). IEEE. DOI: 10.1109/ICICS.2003.1292741

[136] Metallinou, A., Yang, Z., Lee, C. chun, Busso, C., Carnicke, S., & Narayanan, S. (2016). The USC CreativeIT database of multimodal dyadic interactions: from speech and full body motion capture to continuous emotional annotations. *Language resources and evaluation*, *50*(3), 497–521. DOI:10.1007/s10579-015-9300-0

[137] Yang, J. D., Tam, K. W., Huang, T. W., Huang, S. W., Liou, T. H., & Chen, H. C. (2017). Intermittent cervical traction for treating neck pain. *Spine*, *42*(13), 959–965. DOI:10.1097/BRS.0000000000001948

[138] Navas, E., Hernáez, I., & Luengo, I. (2006). An objective and subjective study of the role of semantics and prosodic features in building corpora for emotional TTS. *IEEE transactions on audio, speech and language processing*, *14*(4), 1117–1127. DOI:10.1109/TASL.2006.876121

[139] Puk, K. M., Wang, S., Rosenberger, J., Gandy, K. C., Harris, H. N., Peng, Y. B., … Chiao, J. C. (2022). Emotion recognition and EEG analysis using ADMM-based sparse group lasso. *IEEE transactions on affective computing*, *13*(1), 199–210. DOI:10.1109/TAFFC.2019.2943551

[140] Jiang, M., Dong, J., Ma, D., Sun, J., He, J., & Lang, L. (2022). *Inception spatial temporal graph convolutional networks for skeleton-based action recognition* [presentation]. Proceedings - 2022 international symposium on control engineering and robotics, ISCER 2022 (Vol. 32, pp. 208–213). DOI: 10.1109/ISCER55570.2022.00042

[141] Cui, H., Liu, A., Zhang, X., Chen, X., Wang, K., & Chen, X. (2020). EEG-based emotion recognition using an end-to-end regional-asymmetric convolutional neural network. *Knowledge-based systems*, *205*, 106243. DOI:10.1016/j.knosys.2020.106243

[142] Hassan, M. M., Alam, M. G. R., Uddin, M. Z., Huda, S., Almogren, A., & Fortino, G. (2019). Human emotion recognition using deep belief network architecture. *Information fusion*, *51*, 10–18. DOI:10.1016/j.inffus.2018.10.009

[143] Abdu, S. A., Yousef, A. H., & Salem, A. (2021). Multimodal video sentiment analysis using deep learning approaches, a survey. *Information fusion*, *76*, 204–226. DOI:10.1016/j.inffus.2021.06.003

[144] Petridis, S., & Pantic, M. (2010). Audiovisual discrimination between speech and laughter: Why and when visual information might help. *IEEE transactions on multimedia*, *13*(2), 216–234.

[145] Wu, D., Zhang, J., & Zhao, Q. (2020). Multimodal fused emotion recognition about expression-EEG interaction and collaboration using deep learning. *IEEE access*, *8*, 133180–133189. DOI:10.1109/ACCESS.2020.3010311

[146] Ma, J., Zheng, W. L., Tang, H., & Lu, B. L. (2019). *Emotion recognition using multimodal residual lstm network* [presentation]. MM 2019 - proceedings of the 27th acm international conference on multimedia (pp. 176–183). DOI: 10.1145/3343031.3350871

[147] Yin, Z., Zhao, M., Wang, Y., Yang, J., & Zhang, J. (2017). Recognition of emotions using multimodal physiological signals and an ensemble deep learning model. *Computer methods and programs in biomedicine*, *140*, 93–110. DOI:10.1016/j.cmpb.2016.12.005

[148] Fourati, N., & Pelachaud, C. (2015). Multi-level classification of emotional body expression. In *2015 11th IEEE international conference and workshops on automatic face and gesture recognition (FG)* (Vol. 1, pp. 1-8). IEEE. DOI: 10.1109/FG.2015.7163145

[149] Zhen, Q., Huang, D., Drira, H., Amor, B. Ben, Wang, Y., & Daoudi, M. (2019). Magnifying subtle facial motions for effective 4D expression recognition. *IEEE transactions on affective computing*, *10*(4), 524–536. DOI:10.1109/TAFFC.2017.2747553

[150] Milton, A., Sharmy Roy, S., & Tamil Selvi, S. (2013). SVM scheme for speech emotion recognition using MFCC feature. *International journal of computer applications*, *69*(9), 34–39. DOI:10.5120/11872-7667

[151] Pak, A., & Paroubek, P. (2011). *Text representation using dependency tree subgraphs for sentiment analysis* [presentation]. Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics) (Vol. 6637 LNCS, pp. 323–332). DOI: 10.1007/978-3-642-20244-5_31

[152] Xing, F. Z., Cambria, E., & Welsch, R. E. (2018). Intelligent asset allocation via market sentiment views. *IEEE computational intelligence magazine*, *13*(4), 25–34. DOI:10.1109/MCI.2018.2866727

[153] Ali, K., & Hughes, C. E. (2019). *Facial expression recognition using disentangled adversarial learning*. http://arxiv.org/abs/1909.13135

[154] Conneau, A., Schwenk, H., Cun, Y. Le, & Barrault, L. (2017). *Very deep convolutional networks for text classification*. https://doi.org/10.48550/arXiv.1606.01781

[155] Zhang, S., Zhang, S., Huang, T., & Gao, W. (2017). Speech emotion recognition using deep convolutional neural network and discriminant temporal pyramid matching. *IEEE transactions on multimedia*, *20*(6), 1576–1590.

[156] Fu, Y., Wu, X., Li, X., Pan, Z., & Luo, D. (2020). Semantic neighborhood-aware deep facial expression recognition. *IEEE transactions on image processing*, *29*, 6535–6548.

[157] Mishra, A., Tamilselvam, S., Dasgupta, R., Nagar, S., & Dey, K. (2018). *Cognition-cognizant sentiment analysis with multitask subjectivity summarization based on annotators' gaze behavior* [presentation]. Proceedings of the aaai conference on artificial intelligence (Vol. 32). https://ojs.aaai.org/index.php/AAAI/article/view/12068

[158] Tzirakis, P., Zhang, J., & Schuller, B. W. (2018). End-to-end speech emotion recognition using deep neural networks. *ICASSP, IEEE international conference on acoustics, speech and signal processing - proceedings* (Vol. 2018-April, pp. 5089–5093). IEEE. DOI: 10.1109/ICASSP.2018.8462677

[159] Shen, Z., Cheng, J., Hu, X., & Dong, Q. (2019). *Emotion recognition based on multi-view body gestures* [presentation]. Proceedings - international conference on image processing, ICIP (Vol. 2019-September, pp. 3317–3321). DOI: 10.1109/ICIP.2019.8803460

[160] Sahu, S., Gupta, R., & Espy-Wilson, C. (2022). Modeling feature representations for affective speech using generative adversarial networks. *IEEE transactions on affective computing*, *13*(2), 1098–1110. DOI:10.1109/TAFFC.2020.2998118

[161] Yu, J., Zhang, C., Song, Y., & Cai, W. (2021). *ICE-gan: identity-aware and capsule-enhanced gan with graph-based reasoning for micro-expression recognition and synthesis* [presentation]. Proceedings of the international joint conference on neural networks (Vol. 2021, pp. 1–8). DOI: 10.1109/IJCNN52387.2021.9533988

[162] AbdelWahab, M., & Busso, C. (2018). Domain adversarial for acoustic emotion recognition. *IEEE/acm transactions on audio speech and language processing*, *26*(12), 2423–2435. DOI:10.1109/TASLP.2018.2867099

[163] Li, Z., Zhang, Y., Wei, Y., Wu, Y., & Yang, Q. (2017). *End-to-end adversarial memory network for cross-domain sentiment classification* [presentation]. IJCAI international joint conference on artificial intelligence (Vol. 0, pp. 2237–2243). DOI: 10.24963/ijcai.2017/311

[164] Gera, D., & Balasubramanian, S. (2021). Landmark guidance independent spatio-channel attention and complementary context information based facial expression recognition. *Pattern recognition letters*, *145*, 58–66. DOI:10.1016/j.patrec.2021.01.029

[165] Yang, H. C., & Lee, C. C. (2019). An attribute-invariant variational learning for emotion recognition using physiology. *ICASSP 2019-2019 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 1184-1188). IEEE. DOI: 10.1109/ICASSP.2019.8683290

[166] Yang, H., Ciftci, U., & Yin, L. (2018). *Facial expression recognition by de-expression residue learning* [presentation]. Proceedings of the IEEE computer society conference on computer vision and pattern recognition (pp. 2168–2177). DOI: 10.1109/CVPR.2018.00231

[167] Le Ngo, A. C., Phan, R. C. W., & See, J. (2015). Spontaneous subtle expression recognition: Imbalanced databases and solutions. *Computer vision--ACCV 2014: 12th Asian conference on computer vision* (pp. 33-48). Springer International Publishing. DOI: 10.1007/978-3-319-16817-3_3

[168] Bao, F., Neumann, M., & Vu, N. T. (2019). *CycleGAN-based emotion style transfer as data augmentation for speech emotion recognition* [presentation]. Proceedings of the annual conference of the international speech communication association, interspeech (Vol. 2019, pp. 2828–2832). DOI: 10.21437/Interspeech.2019-2293

[169] Jia, X., Ben, X., Yuan, H., Kpalma, K., & Meng, W. (2018). Macro-to-micro transformation model for micro-expression recognition. *Journal of computational science*, *25*, 289–297. DOI:10.1016/j.jocs.2017.03.016

[170] Zhang, H., Su, W., Yu, J., & Wang, Z. (2020). *Weakly supervised local-global relation network for facial expression recognition* [presentation]. IJCAI international joint conference on artificial intelligence (Vol. 2021-January, pp. 1040–1046). DOI: 10.24963/ijcai.2020/145

[171] Schuller, B., Valstar, M., Eyben, F., McKeown, G., Cowie, R., & Pantic, M. (2011). *AVEC 2011 - the first international audio/visual emotion challenge* [presentation]. Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics) (Vol. 6975 LNCS, pp. 415–424). DOI: 10.1007/978-3-642-24571-8_53

[172] Ringeval, F., Cowie, R., Amiriparian, S., Michaud, A., Schuller, B., Kaya, H., … Pantic, M. (2018). *AVEC 2018 workshop and challenge: bipolar disorder and cross-cultural affect recognition* [presentation]. AVEC 2018 - proceedings of the 2018 audio/visual emotion challenge and workshop, co-located with mm 2018 (pp. 3–13). DOI: 10.1145/3266302.3266316

[173] Pei, E., Xia, X., Yang, L., Jiang, D., & Sahli, H. (2016). Deep neural network and switching Kalman filter based continuous affect recognition. *2016 IEEE international conference on multimedia & expo workshops (ICMEW)* (pp. 1-6). IEEE. DOI: 10.1109/ICMEW.2016.7574729

[174] Song, Y., Morency, L. P., & Davis, R. (2013). *Learning a sparse codebook of facial and body microexpressions for emotion recognition* [presentation]. ICMI 2013 - proceedings of the 2013 ACM international conference on multimodal interaction (pp. 237–244). DOI: 10.1145/2522848.2522851

[175] Valstar, M., Gratch, J., Schuller, B., Ringeval, F., Cowie, R., & Pantic, M. (2016). *Summary for avec 2016: depression, mood, and emotion recognition workshop and challenge* [presentation]. Proceedings of the 24th ACM international conference on multimedia (pp. 1483–1484). https://dl.acm.org/doi/abs/10.1145/2964284.2980532

[176] Zhang, Y., Song, D., Li, X., Zhang, P., Wang, P., Rong, L., … Wang, B. (2020). A Quantum-Like multimodal network framework for modeling interaction dynamics in multiparty conversational sentiment analysis. *Information fusion*, *62*, 14–31. DOI:10.1016/j.inffus.2020.04.003

[177] Tzirakis, P., Chen, J., Zafeiriou, S., & Schuller, B. (2021). End-to-end multimodal affect recognition in real-world environments. *Information fusion*, *68*, 46–53. DOI:10.1016/j.inffus.2020.10.011

[178] Li, W., Zhu, L., Shi, Y., Guo, K., & Cambria, E. (2020). User reviews: Sentiment analysis using lexicon integrated two-channel CNN–LSTM family models. *Applied soft computing journal*, *94*, 106435. DOI:10.1016/j.asoc.2020.106435

[179] Camacho, D., Luzón, M. V., & Cambria, E. (2021). New trends and applications in social media analytics. *Future generation computer systems*, *114*, 318-321.

[180] Cambria, E., Havasi, C., & Hussain, A. (2012). *SenticNet 2: a semantic and affective resource for opinion mining and sentiment analysis* [presentation]. Proceedings of the 25th international Florida artificial intelligence research society conference, flairs-25 (pp. 202–207). http://www.sentic.net/senticnet-2.pdf

[181] Cambria, E., Li, Y., Xing, F. Z., Poria, S., & Kwok, K. (2020). *SenticNet 6: ensemble application of symbolic and subsymbolic ai for sentiment analysis* [presentation]. International conference on information and knowledge management, proceedings (pp. 105–114). DOI: 10.1145/3340531.3412003

[182] Ma, Y., Peng, H., & Cambria, E. (2018). *Targeted aspect-based sentiment analysis via embedding commonsense knowledge into an attentive lstm* [presentation]. 32nd AAAI conference on artificial intelligence, aaai 2018 (Vol. 32, pp. 5876–5883). DOI: 10.1609/aaai.v32i1.12048

[183] Khatua, A., Khatua, A., & Cambria, E. (2020). Predicting political sentiments of voters from Twitter in multi-party contexts. *Applied soft computing journal*, *97*, 106743. DOI:10.1016/j.asoc.2020.106743

[184] Picasso, A., Merello, S., Ma, Y., Oneto, L., & Cambria, E. (2019). Technical analysis and sentiment embeddings for market trend prediction. *Expert systems with applications*, *135*, 60–70. DOI:10.1016/j.eswa.2019.06.014

[185] Qian, Y., Zhang, Y., Ma, X., Yu, H., & Peng, L. (2019). EARS: Emotion-aware recommender system based on hybrid information fusion. *Information fusion*, *46*, 141–146. DOI:10.1016/j.inffus.2018.06.004

[186] Xu, D., Tian, Z., Lai, R., Kong, X., Tan, Z., & Shi, W. (2020). Deep learning based emotion analysis of microblog texts. *Information fusion*, *64*, 1–11. DOI:10.1016/j.inffus.2020.06.002

[187] Yang, D., Alsadoon, A., Prasad, P. W. C., Singh, A. K., & Elchouemi, A. (2018). An Emotion recognition model based on facial recognition in virtual learning environment. *Procedia computer science*, *125*, 2–10.

[188] Zuheros, C., Martínez-Cámara, E., Herrera-Viedma, E., & Herrera, F. (2021). Sentiment Analysis based multi-person multi-criteria decision making methodology using natural language processing and deep learning for smarter decision aid. Case study of restaurant choice using TripAdvisor reviews. *Information fusion*, *68*, 22–36. DOI:10.1016/j.inffus.2020.10.019

[189] Kaza, K., Psaltis, A., Stefanidis, K., Apostolakis, K. C., Thermos, S., Dimitropoulos, K., & Daras, P. (2016). *Body motion analysis for emotion recognition in serious games* [presentation]. Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics) (Vol. 9738, pp. 33–42). DOI: 10.1007/978-3-319-40244-4_4

[190] Cai, H., Qu, Z., Li, Z., Zhang, Y., Hu, X., & Hu, B. (2020). Feature-level fusion approaches based on multimodal EEG data for depression recognition. *Information fusion*, *59*, 127–138. DOI:10.1016/j.inffus.2020.01.008

[191] Dong, W., Yang, L., Gravina, R., & Fortino, G. (2021). ANFIS fusion algorithm for eye movement recognition via soft multi-functional electronic skin. *Information fusion*, *71*, 99–108. DOI:10.1016/j.inffus.2021.02.003

[192] Piana Dibris, S., Staglianò Dibris, A., Camurri Dibris, A., & Odone Dibris, F. (2013). *A set of full-body movement features for emotion recognition to help children affected by autism spectrum condition* [presentation]. IDGEI international workshop. (Vol. 23, pp. 1–7). http://www.fdg2013.org/program/workshops/papers/IDGEI2013/idgei2013_4.pdf

[193] Sawada, L. O., Mano, L. Y., Torres Neto, J. R., & Ueyama, J. (2019). A module-based framework to emotion recognition by speech: a case study in clinical simulation. *Journal of ambient intelligence and humanized computing*, 1–10. DOI:10.1007/s12652-019-01280-8

[194] Kanjo, E., Younis, E. M. G., & Ang, C. S. (2019). Deep learning analysis of mobile physiological, environmental and location sensor data for emotion detection. *Information fusion*, *49*, 46–56. DOI:10.1016/j.inffus.2018.09.001

[195] Low, L. S. A., Maddage, N. C., Lech, M., Sheeber, L. B., & Allen, N. B. (2011). Detection of clinical depression in adolescents' speech during family interactions. *IEEE transactions on biomedical engineering*, *58*(3 PART 1), 574–586. DOI:10.1109/TBME.2010.2091640

[196] Schmidt, P., Reiss, A., Duerichen, R., & Van Laerhoven, K. (2018). *Introducing wesad, a multimodal dataset for wearable stress and affect detection* [presentation]. ICMI 2018 - proceedings of the 2018 international conference on multimodal interaction (pp. 400–408). DOI: 10.1145/3242969.3242985

[197] Kim, D. H., Lee, M. K., Choi, D. Y., & Song, B. C. (2017). *Multi-modal emotion recognition using semi-supervised learning and multiple neural networks in the wild* [presentation]. ICMI 2017 - proceedings of the 19th acm international conference on multimodal interaction (Vol. 2017, pp. 529–535). DOI: 10.1145/3136755.3143005

[198] Tsiami, A., Koutras, P., Efthymiou, N., Filntisis, P. P., Potamianos, G., & Maragos, P. (2018). Multi3: Multi-sensory perception system for multi-modal child interaction with multiple robots. *2018 IEEE international conference on robotics and automation (ICRA)* (pp. 4585-4592). IEEE. DOI: 10.1109/ICRA.2018.8461210